Eva Karina Dos Santos De Oliveira

MODELAÇÃO DO PREÇO DE IMÓVEIS PARA HABITAÇÃO NO CONCELHO DE LISBOA

DOCUMENTO PROVISÓRIO

Universidade de Aveiro Departamento de Matemática
2015

Eva Karina Dos Santos De Oliveira

MODELAÇÃO DO PREÇO DE IMÓVEIS PARA HABITAÇÃO NO CONCELHO DE LISBOA

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Matemática e Aplicações, realizada sob a orientação científica do Doutor Luís Silva, Professor auxiliar convidado do Departamento de Matemática da Universidade de Aveiro.

Dedico este trabalho aos meus pais.

o júri / the jury

presidente / president

Professor Doutor Agostinho Miguel Mendes Agra
Professor auxiliar da Universidade de Aveiro

vogais / examiners committee

Professora Doutora Sandra Cristina de Faria Ramos
Professora adjunta do Instituto Superior de Engenharia do Porto

Professor Doutor Luís Miguel Almeida da Silva
Professor auxiliar convidado da Universidade de Aveiro

agradecimentos /
acknowledgements

Estou grata à empresa Ubiwhere pela oportunidade de estágio.
Ao Dr. Luís Silva, meu orientador, pelo apoio e disponibilidade demonstrada ao longo do projeto, pelas correções, sugestões e pela orientação dada.
Aos meus pais, pela paciência e compreensão.
Ao meu irmão, Miguel, pela motivação.
Aos meus padrinhos, José e Ângela, pela presença de todos os dias.
Ao Juan pelo apoio incondicional.
Por último, a todos os meus amigos que me acompanharam nesta jornada e a tornaram mais fácil e feliz!

Palavras Chave regressão linear múltipla, avaliação de imóveis, preço de venda de imóveis, mercado imobiliário

Resumo O presente trabalho tem como principais objetivos reconhecer os atributos relevantes para a formação do preço de venda de imóveis na cidade de Lisboa, bem como construir um modelo de previsão de preços de imóveis para a mesma cidade. Este trabalho foi proposto pela empresa Ubiwhere cujo resultado servirá de apoio à construção de uma ferramenta de avaliação de imóveis que integrará a plataforma imobiliária Livin'X que está a ser desenvolvida pela empresa. Para atingir os objetivos propostos foram elaborados modelos de previsão com base no modelo de regressão linear múltipla e utilizada a análise hedónica para identificação dos atributos estruturais e de localização da habitação mais relevantes na formação do seu preço. Foi utilizada a técnica de validação cruzada para mais assertivamente selecionar os modelos mais explicativos e para esses foram analisados os pressupostos do modelo de regressão linear aplicado. A proposta de solução ao problema contém informação sobre as variáveis relevantes na construção do preço de imóveis na cidade de Lisboa, bem como modelos de previsão de preços em diferentes perspetivas.

Keywords multiple linear regression, real estate market, real estate price

Abstract

The main goals of this study to recognize the relevant attributes to the formation of real estate selling price in the city of Lisbon, as well as to establish a real estate price forecasting model to this city. This study was proposed by the Ubiwhere company whose outcome will serve as a support to the construction of an assessment tool for real estate which will be integrated in the Livin'X platform that is being developed by this company. To achieve the proposed objectives, predictive models based on multiple linear regression were developed and the hedonic analysis approach was used to identify both structural and location attributes that are most relevant to the formation of the real estate selling price. We used the cross-validation technique for a more assertively selection of the most explanatory models and for these, the assumptions of multiple linear regression were analyzed. The proposed solution to this problem contains information on the relevant variables in the process of determination for the price of real state in the city of Lisbon, as well as price forecasting models in different perspectives.

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

O mercado imobiliário é altamente concorrencial, exigente e dinâmico, pelo que se torna cada vez mais importante identificar os fatores que influenciam a variabilidade de preços dos imóveis de forma a compreender o mercado atual e a disponibilizar respostas adequadas e oportunas.

A avaliação de imóveis visa estimar um valor de mercado de um imóvel de acordo com as suas características, no entanto cada imóvel é único, dado que não existem dois imóveis exatamente iguais, particularidade que dificulta a sua avaliação.

Neste sentido, este trabalho contribui com uma proposta de um modelo de avaliação de imóveis para habitação na cidade de Lisboa e identificação dos atributos que influenciam a formação do seu preço, de acordo com os dados disponíveis. Os resultados serão utilizados no desenvolvimento de uma ferramenta de avaliação de imóveis integrada numa plataforma imobiliária.

1.1 Apresentações da empresa

A *Ubiwhere* é uma empresa focada no desenvolvimento e investigação de tecnologias de ponta, para conceber a tecnologia mais avançada e criar propriedade intelectual de grande valor, quer a nível interno quer para os seus clientes. As suas atividades envolvem a prestação de serviços nas áreas da consultoria e desenvolvimento de *software* e investigação na área da computação úbiqua.

O trabalho elaborado para a empresa enquadra-se no desenvolvimento de uma plataforma que acompanha a tendência das SmartCities. Esta plataforma tem o nome de Livin'X e permite aos seus utilizadores procurar ou publicar imóveis de habitação, para venda ou arrendamento, com funcionalidades muito próprias e inovadoras. Neste sentido, a plataforma Livin'X ajuda os seus utilizadores a procurar o melhor sítio para viver, tendo em consideração não só as características estruturais pretendidas para o imóvel, como também a sua envolvente e as preferências e interesses do seu utilizador.

Neste contexto a empresa está interessada em identificar os atributos valorativos dos imóveis para habitação e construir uma ferramenta tecnológica que permita prever o preço de

um imóvel em função dos interesses e necessidades do utilizador da plataforma Livin'X, que tenciona comprar casa na cidade de Lisboa.

1.2 Mercado da habitação

A habitação, na sociedade, é um objeto que tem como função principal a de abrigo para a família e é ainda, para cada indivíduo que a ocupa, um elemento fundamental para a construção da sua personalidade, de integração social e de socialização.

O processo de seleção de informação relevante para a avaliação de um imóvel é complicado, pois estando inserida num espaço de consumo, de produção de bens e serviços, lazer e comunicação, a habitação é valorizada não só pelos seus atributos estruturais, mas também pelo conjunto de oportunidades que lhe estão associadas, como a distância ao emprego, à escola das crianças, a serviços públicos e comércio, espaços verdes, níveis de conforto e segurança, etc. Estas características são valorizadas de forma diferente por diferentes indivíduos, o que dificulta a sua avaliação. Para além dos aspetos mencionados, existem outros atributos da habitação que lhe são característicos, como a heterogeneidade (não há dois imóveis iguais), a imobilidade e a durabilidade (quer a nível físico, quer a nível de investimento) e ainda aspetos económicos de nível nacional como a regulamentação e impostos, difíceis de incorporar num modelo de avaliação.

Avaliar um imóvel é estimar um valor em função das características e oportunidades a ele associadas, determinadas condições do mercado e o fim a que se destina. A fiabilidade de uma avaliação depende não só da competência do avaliador, mas também da disponibilidade e recolha de dados pertinentes. Por lhe estar associado tantos e tão variados fatores, a avaliação imobiliária é uma atividade multidisciplinar que requer um amplo leque de conhecimentos. Como tal, quando se pretende criar um método de avaliação de preços de imóveis é conveniente que sejam articulados conhecimentos dos profissionais das diversas áreas envolvidas como: profissionais imobiliários, que pela sua proximidade ao mercado conhecem o comportamento da oferta e da procura, os preços praticados e as tendências; profissionais de planeamento urbano, que são capazes de avaliar a zona em que estão inseridos os imóveis bem como o seu potencial; construtores do ramo imobiliário, que são conhecedores do custo da construção dos imóveis, entre outros.

Métodos de avaliação imobiliária

A avaliação de uma propriedade imobiliária é importante pois é utilizada em vários ramos de atividade, não sendo utilizada apenas para a compra ou venda mas também para outorgar financiamentos, estudos económicos e financeiros de projetos de investimento, o cálculo de indemnização por expropriação, cálculo de prémios de seguro, determinação do valor de impostos, etc.

Os métodos de avaliação a seguir apresentados têm aderência à realidade portuguesa pois são métodos utilizados no crédito hipotecário e na avaliação das propriedades dos fundos de investimento imobiliário (Tavares, 2011).

Método Comparativo

Este método consiste em estimar o preço de um imóvel comparando-o com propriedades semelhantes, isto é, o imóvel a avaliar é comparado, caso existam, a imóveis com as mesmas características quer estruturais, quer de localização.

Contudo, não existindo dois imóveis exatamente iguais, torna-se necessário realizar ajustamentos ao preço do imóvel alvo da avaliação. Nesses ajustamentos são tidos em conta quer as caraterísticas intrínsecas ao imóvel como também condições de venda, de financiamento e de mercado. Os ajustamentos são algo subjetivos, pelo que devem ser realizados por profissionais com experiência na avaliação de imóveis e conhecedores das tendências e condições do mercado imobiliário.

Método de Rendimento

O método do rendimento considera que o valor de um ativo é função dos fluxos de rendimento que este é capaz de gerar.

Para estimar o rendimento que um imóvel é capaz de gerar devem ser tidas em conta:

- O valor da renda de um imóvel ou o valor da renda de imóveis semelhantes
- As despesas operacionais e de investimento
- O histórico de arrendamentos do imóvel (ou imóveis semelhantes) de forma a incluir na análise os períodos de tempo que o imóvel poderá estar desocupado sem que produza rendimento.

Método do Custo

O método do custo considera que o valor de um imóvel resulta da soma dos custos de produção com o lucro do vendedor e/ou promotor. Baseia-se na ideia de que o preço de um produto é a soma dos custos de cada um dos seus componentes, ou seja, o valor do imóvel obtém-se adicionando ao valor do terreno e respetivos encargos com a sua aquisição, o custo da construção eventualmente depreciado em função da obsolescência física, funcional, ambiental e económica, e apreciado em função de singularidades arquitetónicas, históricas ou outras verificadas. Baseia-se também na ideia de que um comprador não pagaria menos se comprasse o terreno e construísse o edifício.

Este método é geralmente aplicado na avaliação de imóveis raramente transacionados e não vocacionados para o lucro como hospitais, escolas, prisões, bibliotecas, museus, castelos, entre outros; avaliação de edifícios antigos; avaliação de construções para efeitos de fixação

de prémios de seguro, indemnizações e tributações fiscais e outros; avaliação de obras para reabilitação.

Método de Avaliação Residual

O método do valor residual é um caso particular do método do custo pois considera no seu processo de cálculo todos os custos e receitas envolvidos na execução do empreendimento imobiliário. Este método aplica-se na estimativa do valor de bens imobiliários com um valor potencial, ou seja, cujo valor poderá ser substancialmente superior se forem investidos capitais de modo a promover a sua alteração ou ampliação.

Este método é utilizado na avaliação de propriedades com potencial de desenvolvimento, como terrenos baldios ou imóveis com potencial de serem renovados ou verem o seu uso alterado para um fim mais lucrativo, como no caso de conversão de uma fábrica antiga num conjunto de apartamentos.

Os princípios em que se baseia este método são o de melhor uso e do valor residual.

Modelo Hedónico

Os modelos hedónicos baseiam-se num modelo matemático que apresenta como variáveis independentes os atributos dos imóveis e como resposta o preço dos imóveis.

O objetivo deste método é encontrar os atributos que explicam a formação do preço dos imóveis e determinar a importância relativa de cada um deles. Trata-se de uma metodologia baseada em modelos estatísticos lineares e não lineares para determinar a relevância parcial de cada um dos atributos envolvidos no modelo.

O modelo hedónico é aplicado quando os dados sobre os imóveis são heterogéneos, utilizando informações concretas referentes a um determinado número de imóveis e permitindo estimar o preço de outros imóveis não envolvidos mediante a aplicação do modelo matemático obtido.

A aplicação deste método traz algumas desvantagens: é necessário ter em atenção a problemas de multicolinearidade de forma a não incluir variáveis não adequadas ou desnecessárias, é difícil caracterizar/atribuir valor a variáveis relativas a atributos de localização e vizinhança; é necessário um elevado número de imóveis para assegurar um boa precisão, entre outros (González e Formoso, 2000) (Tavares, 2011).

A escolha do método a utilizar depende dos objetivos da avaliação. Em Portugal os métodos do custo e comparativo são os mais utilizados, no entanto têm surgido recentemente propostas de aplicação dos modelos hedónicos tais como as sugeridas por Batista (2010), Catalão (2010), Guedes (2011), Marques (2012) e Vigas (2013). Estas propostas têm por base modelos de regressão linear múltipla que surge em alguns trabalhos combinada com análise fatorial e/ou análise em componentes principais. Também existem trabalhos que aplicam os modelos hedónicos baseados em modelos de regressão não linear, nomeadamente a utilização

de redes neuronais artificiais tais como os propostos por Couto (2007) e Moreira, Silva e Fernandes (2010).

Neste trabalho, o modelo de avaliação imobiliária sugerido é baseado nos modelos hedónicos pela sua forte componente estatística que se enquadra no âmbito do mestrado e também porque é o método de avaliação que identifica as variáveis relevantes para a formação do preço dos imóveis, um dos principais propósitos deste trabalho.

Identificação dos Atributos Relevantes

Para aplicar o modelo hedónico é importante que sejam incorporados os atributos relevantes para a explicação da variabilidade do preço dos imóveis de habitação. Na tentativa de encontrar esses atributos de forma assertiva foram consultados estudos sobre a matéria.

Quanto aos atributos estruturais da habitação, aqueles que são importantes incluir na sua avaliação são a tipologia, as áreas útil e/ou bruta, a natureza, a idade do imóvel e o estado de conservação. Outros atributos que são frequentemente avaliados são a existência de ar condicionado, aquecimento central, arrecadação, aspiração, churrasqueira, despensa, estacionamento, garagem, hidromassagem, *jacuzzi*, jardim, cozinha equipada, *kitchenette*, lareira, lavandaria, marquise, pátio, segurança, sótão, terraço/varanda, elevador, piscina, roupeiros, suite, gás canalizado, acessos para pessoas com mobilidade reduzida, cave, se o imóvel está mobilado, se é duplex, se está localizado num condominio fechado, se é uma moradia isolada ou um apartamento no rés-do-chão ou último andar, se é um imóvel da banca, entre outros.

É óbvio que os atributos estruturais de um imóvel sejam relevantes na construção do seu preço, no entanto, alguns estudos realçam a importância da localização do imóvel na formação do seu preço (Kiel e Zabel, 2008). Numa tentativa de encontrar os atributos relativos à localização do imóvel relevantes para a sua avaliação, inúmeros estudos foram elaborados tendo sido obtidas algumas conclusões interessantes. Tavares (2011) no seu trabalho "Avaliação Imobiliária - Entre a Ciência da Avaliação e a Arte da Apreciação" faz uma análise desses estudos produzindo uma compilação dos atributos relativos à localização dos imóveis apontados como relevantes. A estes atributos o autor dá o nome de "externalidades" e classifica-as em positivas, negativas e nas que podem ser vistas como positivas ou negativas. Das externalidades positivas, i.e, externalidades que acrescentam valor ao imóvel, o autor destaca a proximidade às escolas, lojas, transportes públicos, áreas verdes, vista de mar e/ou rio, existência de ajardinamentos nos passeios, parques de recreio, temperaturas amenas, qualidade paisagística e planeamento urbano e imóveis construídos segundo os princípios da arquitetura bioclimática. Quanto às externalidades negativas, i.e, aquelas que desvalorizam o imóvel, o autor destaca a proximidade a auto-estradas, caixas de correio e escadas, exposição a vistas e odores indesejáveis, locais com maiores quantidades de chuva, temperaturas de canícula no verão e maiores índices de humidade, proximidade a aterros sanitários, centrais de carvão, refinarias químicas, centrais nucleares e turbinas eólicas, locais propícios à ocorrência

de inundações, terramotos, furacões e tornados. As externalidades que podem ser entendidas como positivas ou negativas, dependendo das preferências do comprador, são a qualidade da água e do ar, o rendimento médio e taxa de desemprego da vizinhança.

A partir da bibliografia estudada foi extraída informação sobre as características mais importantes ou preferidas pelos indivíduos que adquirem imóveis para habitação e o seu impacto na formação do preço, com vista a ser utilizada na recolha dos dados. Como a recolha de dados ficou a cargo da empresa, não foi possível selecionar os atributos a coletar. No entanto foi feito um esforço no sentido de completar a base de dados cedida pela empresa com dados obtidos posteriormente baseados nesta pesquisa.

1.3 Metodologia de investigação

De forma a compreender melhor a problemática da avaliação imobiliária foi realizado um estudo de investigação sobre o mercado imobiliário em Portugal, os modelos e teorias aplicados com o fim de avaliar imóveis e identificar as características mais relevantes para a sua avaliação.

Foi efetuada uma pesquisa sobre o *software* apropriado à análise pretendida, da qual resultou a escolha do IBM SPSS Statistics para realizar todas as análises estatísticas.

A recolha dos dados foi feita pela empresa. Após a cedência da base de dados inicial foram inseridos dados complementares considerados pertinentes após a pesquisa bibliográfica. Seguiu-se uma "limpeza" da base de dados obtida e a sua preparação de forma a ser compreendida pelo *software*. Realizou-se também uma análise descritiva aos dados de forma a compreender e sumariar a informação neles contida.

Após esta análise preliminar, foi aplicado o modelo de regressão linear múltipla, com o método dos mínimos quadrados, validado através da técnica de validação cruzada. Este modelo será apresentado com mais pormenor no capítulo 2 deste relatório.

Da aplicação da regressão linear múltipla resultaram cinco modelos: o primeiro avalia os imóveis para habitação de uma forma geral, enquanto os seguintes avaliam apartamentos e moradias separadamente e ainda dois modelos que avaliam imóveis de classe média/baixa e classe alta (estas definições serão descritas mais adiante). A partir destes modelos é possível identificar os atributos relevantes na formação do preço dos imóveis e fazer previsão do mesmo.

1.4 Organização do relatório

Este relatório está dividido em 4 capítulos.

No capítulo 2 apresenta-se o modelo de regressão linear múltipla, os seus pressupostos, técnicas de seleção de variáveis, validação do modelo e análise de observações influentes.

No capítulo 3 é apresentado o problema e uma proposta de solução. Este capítulo está dividido em três partes. Na primeira é descrito o processo de tratamento e limpeza da base

de dados; na segunda é apresentada uma análise descritiva dos dados; na terceira é descrito todo o processo experimental até à obtenção da solução proposta, com base no modelo de regressão linear múltipla, e a sua discussão.

No último capítulo são apresentadas as conclusões e sugestões para trabalho futuro.

Capítulo 2

Regressão Linear

A análise de regressão é uma técnica estatística utilizada para estudar e modelar relações entre uma variável dependente e uma ou mais variáveis independentes. Tem inúmeras aplicações em diversas áreas e é utilizada frequentemente para sumariar dados, estimar parâmetros, fazer previsão e controlo.

Aplicada à avaliação imobiliária, a regressão linear pode ser utilizada para estimar a contribuição relativa das variáveis (atributos do imóvel) na construção do seu preço e para fazer previsão do valor ou um intervalo de valores admissível para o preço do imóvel em causa.

2.1 O modelo de regressão linear

Um modelo de regressão linear descreve a relação entre uma variável quantitativa dependente (ou resposta), y, e um conjunto de k variáveis quantitativas regressoras, $x_j; j = 1, ..., k$ da seguinte forma:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \epsilon. \tag{2.1}$$

A equação (2.1) descreve um hiperplano no espaço k-dimensional dos regressores x_j. O parâmetro β_0 é a ordenada na origem, designada constante daqui em diante. Caso o vetor nulo pertença ao domínio dos regressores, β_0 pode ser interpretado como a média de y nesse ponto, caso contrário não existe interpretação prática para β_0. Os parâmetros $\beta_j, j = 1, ..., k$ representam a variação de y por unidade de variação de x_j, quando os restantes regressores $x_i (i \neq j)$ se mantêm constantes; por esta razão estes parâmetros são também chamados de coeficientes parciais de regressão. O erro aleatório ϵ reflete os erros de medição e a variação natural em y. Neste modelo y é variável aleatória e $x_j, j = 1, ..., k$ são não aleatórias. Este modelo é linear relativamente aos parâmetros $\beta_0, \beta_1, ..., \beta_k$.

Assume-se assim que o modelo descreve a i-ésima resposta y_i por

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \epsilon_i, \quad i = 1, ..., n. \tag{2.2}$$

Devido ao elevado número de parâmetros, torna-se conveniente expressar as operações matemáticas utilizando notação matricial. Assim, as equações (2.2) podem ser representadas por:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.3}$$

onde

$$\boldsymbol{y}^T = \begin{bmatrix} y_1 & y_2 & \dots & y_n \end{bmatrix}^T$$

é o vetor das observações;

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}$$

é a matriz das n observações de x_j, onde a primeira coluna foi preenchida com 1's de forma a ser possível obter a estimativa para β_0;

$$\boldsymbol{\beta}^T = \begin{bmatrix} \beta_0 & \beta_1 & \dots & \beta_k \end{bmatrix}^T$$

é o vetor dos coeficientes de regressão;

$$\boldsymbol{\epsilon}^T = \begin{bmatrix} \epsilon_1 & \epsilon_2 & \dots & \epsilon_n \end{bmatrix}^T$$

é o vetor dos erros aleatórios. Estes erros são independentes, têm valor médio nulo e variância constante, σ^2. Assim a equação de regressão pode ser escrita da seguinte forma (Montgomery, Peck e Vining, 2006):

$$E[y|\boldsymbol{X}] = \boldsymbol{X}\boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k, \tag{2.4}$$

com

$$V[y|\boldsymbol{X}] = Var[\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \epsilon] = \sigma^2. \tag{2.5}$$

já que a variância de y não depende de $\boldsymbol{X}$ e, como os erros são independentes, as respostas também são não correlacionadas.

2.2 Estimação dos parâmetros do modelo pelos mínimos quadrados

Estimação dos β_j

Como a totalidade da população sob estudo não está, normalmente, acessível, os parâmetros do modelo são estimados a partir de uma amostra representativa. Assim, pretende-se estimar os parâmetros da equação (2.4) utilizando os estimadores apropriados, $\widehat{\beta}_j$, que produzem as estimativas amostrais dos verdadeiros parâmetros da população.

Um método frequentemente utilizado é o método dos mínimos quadrados. Neste método as estimativas dos coeficientes de regressão são calculadas de forma a que a soma dos quadrados dos desvios seja mínima. Assim, o problema de estimação dos β_j resume-se ao problema de minimização da função

$$S(\beta_0, \beta_1, ..., \beta_k) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2 \tag{2.6}$$

em relação a $\beta_0, \beta_1, ..., \beta_k$. Logo, os estimadores dos mínimos quadrados de $\beta_0, \beta_1, ..., \beta_k$ devem satisfazer

$$\left. \frac{\partial S}{\partial \beta_0} \right|_{\widehat{\beta}_0, \widehat{\beta}_1, ..., \widehat{\beta}_k} = -2 \sum_{i=1}^{n} \left(y_i - \widehat{\beta}_0 - \sum_{j=1}^{k} \widehat{\beta}_j x_{ij} \right) = 0, \tag{2.7}$$

$$\left. \frac{\partial S}{\partial \beta_j} \right|_{\widehat{\beta}_0, \widehat{\beta}_1, ..., \widehat{\beta}_k} = -2 \sum_{i=1}^{n} \left(y_i - \widehat{\beta}_0 - \sum_{j=1}^{k} \widehat{\beta}_j x_{ij} \right) x_{ij} = 0, j = 1, 2, ..., k. \tag{2.8}$$

Simplificando as equações (2.7) e (2.8) obtêm-se as seguintes equações normais das estimativas dos mínimos quadrados

$$\begin{aligned}
n\widehat{\beta}_0 + \widehat{\beta}_1 \sum_{i=1}^{n} x_{i1} + \widehat{\beta}_2 \sum_{i=1}^{n} x_{i2} + ... + \widehat{\beta}_k \sum_{i=1}^{n} x_{ik} &= \sum_{i=1}^{n} y_i \\
\widehat{\beta}_0 \sum_{i=1}^{n} x_{i1} + \widehat{\beta}_1 \sum_{i=1}^{n} x_{i1}^2 + \widehat{\beta}_2 \sum_{i=1}^{n} x_{i1} x_{i2} + ... + \widehat{\beta}_k \sum_{i=1}^{n} x_{i1} x_{ik} &= \sum_{i=1}^{n} x_{i1} y_i \\
\vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad\qquad\qquad \vdots \qquad\quad \vdots & \\
\widehat{\beta}_0 \sum_{i=1}^{n} x_{ik} + \widehat{\beta}_1 \sum_{i=1}^{n} x_{ik} x_{i1} + \widehat{\beta}_2 \sum_{i=1}^{n} x_{ik} x_{i2} + ... + \widehat{\beta}_k \sum_{i=1}^{n} x_{ik}^2 &= \sum_{i=1}^{n} x_{ik} y_i
\end{aligned} \tag{2.9}$$

Da solução destas $k+1$ equações obtêm-se os estimadores dos mínimos quadrados para $\beta_0, \beta_1, ..., \beta_k$.

Em termos matriciais, a equação (2.6) pode ser representada por

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \epsilon_i^2 = \boldsymbol{\epsilon}^T \boldsymbol{\epsilon} = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{y}^T \boldsymbol{y} - 2\boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{y} + \boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{\beta}. \tag{2.10}$$

e os estimadores dos mínimos quadrados para $\boldsymbol{\beta}$ devem satisfazer

$$\left. \frac{\partial S}{\partial \boldsymbol{\beta}} \right|_{\widehat{\boldsymbol{\beta}}} = 0 \Longleftrightarrow -2\boldsymbol{X}^T \boldsymbol{y} + 2\boldsymbol{X}^T \boldsymbol{X} \widehat{\boldsymbol{\beta}} = 0 \Longleftrightarrow \boldsymbol{X}^T \boldsymbol{X} \widehat{\boldsymbol{\beta}} = \boldsymbol{X}^T \boldsymbol{y}. \tag{2.11}$$

Da equação (2.11) obtém-se o estimador dos mínimos quadrados de $\boldsymbol{\beta}$

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}, \tag{2.12}$$

desde que a matriz $(\boldsymbol{X}^T \boldsymbol{X})^{-1}$ exista. A matriz $(\boldsymbol{X}^T \boldsymbol{X})^{-1}$ existe sempre que as variáveis regressoras sejam linearmente independentes (Montgomery, Peck e Vining, 2006).

Propriedades do estimador $\widehat{\boldsymbol{\beta}}$

O estimador do mínimos quadrados $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ é uma combinação linear das variáveis regressoras e tem distribuição normal multivariada (Montgomery, Peck e Vining, 2006). Como a matriz $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I}$ vem que

$$\begin{aligned} E[\widehat{\boldsymbol{\beta}}] =& E[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}] = E[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon})] \\ =& E[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon})] = \boldsymbol{\beta}, \end{aligned} \tag{2.13}$$

donde se conclui que $\widehat{\boldsymbol{\beta}}$ é um estimador centrado de $\boldsymbol{\beta}$.

O estimador $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ é linear (Murteira et al., 2010) em $\boldsymbol{y}$, já que a matriz $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ é fixa (não aleatória) e a matriz de variâncias-covariâncias de $\widehat{\boldsymbol{\beta}}$, $\Sigma_{\widehat{\beta}}$, é dada por (Montgomery, Peck e Vining, 2006)

$$\Sigma_{\widehat{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{I}\sigma^2 = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}. \tag{2.14}$$

Tomando $\boldsymbol{C} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}$, temos $Var[\widehat{\beta}_j] = \sigma^2 C_{jj}$ e $Cov[\widehat{\beta}_i, \widehat{\beta}_j] = \sigma^2 C_{ij}$.

Se, para além dos erros serem independentes com valor médio nulo e variância constante, se verificar que os erros são normalmente distribuídos, prova-se que o estimador dos mínimos quadrados coincide com o estimador da máxima verosimilhança que é o estimador não enviesado de variância mínima de $\boldsymbol{\beta}$. Nestas condições, $\widehat{\boldsymbol{\beta}}$ é o estimador mais eficiente de $\boldsymbol{\beta}$, na classe dos estimadores lineares não enviesados (Montgomery, Peck e Vining, 2006).

Coeficientes de regressão *standardizados*

As unidades das variáveis regressoras são, em geral, diferentes pelo que o valor dos seus parâmetros associados não pode ser interpretado diretamente como uma medida de contribuição de cada regressor para a explicação da variação da variável resposta. Por essa razão torna-se útil *standardizar* os coeficientes de regressão de forma a obter coeficientes de regressão adimensionais. Os coeficientes de regressão podem ser *standardizados* pela fórmula (Marôco, 2010):

$$\widehat{\beta}'_j = \widehat{\beta}_j \left(\frac{S_{x_j}}{S_y} \right), \tag{2.15}$$

onde S_{x_j} e S_y representam os desvios padrão amostrais das variáveis x_j e y, respetivamente.

Estes coeficientes dão a taxa de variação em unidades de desvio padrão para y por cada variação de uma unidade de desvio padrão para x_j, mantendo constantes as restantes variáveis regressoras $x_i (i \neq j)$ (Field, 2009). A partir destes coeficientes é possível comparar a importância relativa dos regressores.

Estimação de σ^2

Para além das estimativas dos coeficientes de regressão, é também necessário estimar a variância das variáveis residuais, σ^2, para que seja possível fazer testes de hipóteses e construir

intervalos de confiança pertinentes para o modelo de regressão.
Uma estimativa para σ^2 pode ser obtida através da variação residual (VR),

$$VR = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n} e_i^2. \tag{2.16}$$

Prova-se que $E[VR] = (n-k-1)\sigma^2$ (Montgomery, Peck e Vining, 2006), pelo que um estimador não enviesado de σ é

$$\widehat{\sigma}^2 = \frac{VR}{n-k-1} = QME, \tag{2.17}$$

onde QME é o quadrado médio dos erros. Como $\widehat{\sigma}^2$ é calculado utilizando a variação residual do modelo, diz-se que $\widehat{\sigma}^2$ é condicionado pelo modelo.

Estimados todos os parâmetros obtém-se o modelo ajustado correspondente à equação (2.4):

$$E[\widehat{y}|\boldsymbol{X}] = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \ldots + \widehat{\beta}_k x_k, \tag{2.18}$$

onde

$$V[\widehat{y}|\boldsymbol{X}] = \widehat{\sigma}^2 = QME. \tag{2.19}$$

As diferenças entre os valores observados da variável dependente, y_i, e os valores correspondentes ajustados pela equação de regressão (2.18), $\widehat{y}_i$, dizem-se erros do modelo. Estes podem ser representados por

$$\boldsymbol{e} = \boldsymbol{y} - \widehat{\boldsymbol{y}}. \tag{2.20}$$

Estes erros têm as seguintes propriedades (Murteira et al., 2010):

(a) A soma dos erros é igual a zero: $\sum_{i=1}^{n} e_i = 0$;

(b) A soma dos produtos das observações de cada regressor pelo respetivo erro é igual a zero: $\sum_{i=1}^{n} x_{ij} e_i = 0, j = 1, 2, ..., k$;

(c) A soma dos produtos dos valores ajustados pelo respetivo erro é igual a zero: $\sum_{i=1}^{n} \widehat{y}_i e_i = 0$;

(d) A soma dos quadrados das observações da variável resposta é igual à soma dos quadrados dos respetivos valores ajustados mais a soma dos quadrados dos erros: $\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} \widehat{y}_i^2 + \sum_{i=1}^{n} e_i^2$.

2.3 Testes de hipóteses

Uma vez estimados os parâmetros do modelo é possível avaliar a significância das variáveis independentes sobre a variável dependente realizando testes de hipóteses. Estes testes requerem que os erros aleatórios sejam independentes e normalmente distribuídos com média nula e variância constante (Montgomery, Peck e Vining, 2006).

Teste de significância global

O teste de significância global da regressão testa se existe efetivamente uma relação linear entre a variável resposta, y, e as variáveis regressoras, $x_1, x_2, ..., x_k$, isto é, testa se o modelo ajustado é significativo. A formalização deste teste de hipóteses é a seguinte:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0 \qquad vs. \qquad H_1 : \beta_j \neq 0 \text{ para algum } j.$$

A rejeição da hipótese nula a favor da alternativa leva a concluir que pelo menos um dos regressores possuí um efeito significativo sobre a variável resposta. Para testar estas hipóteses, a variação total (VT) é dividida em variação explicada pelo modelo de regressão (VE) e variação residual (VR):

$$VT = VE + VR.$$

Se a variação explicada é significativamente maior do que a variação residual, pode concluir-se que o modelo ajustado é significativo. A estatística de teste é dada por (Marôco, 2010)

$$F = \frac{\frac{VE}{k}}{\frac{VR}{(n-k-1)}} = \frac{QMR}{QME},$$

onde QMR representa o quadrado médio da regressão. A estatística F segue uma distribuição $F_{k,n-k-1}$.

Teste individual aos coeficientes de regressão

Uma vez confirmado que o modelo ajustado é significativo, a questão que se levanta é a de saber qual ou quais dos regressores possuem uma relação linear significativa com a variável resposta. As hipóteses para testar a significância de cada coeficiente de regressão β_j para $j = 0, 1, ...k$, são:

$$H_0 : \beta_j = 0 \qquad vs. \qquad H_1 : \beta_j \neq 0 \;\; j = 0, 1, ..., k.$$

Se a hipótese nula é rejeitada a favor da alternativa, conclui-se que o regressor x_j pode ser eliminado do modelo. A estatística de teste, sob H_0, é (Montgomery, Peck e Vining, 2006):

$$T = \frac{\widehat{\beta}_j}{\sqrt{\widehat{\sigma}^2 C_{jj}}}, \tag{2.21}$$

onde C_{jj} é o elemento j da diagonal da matriz $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ correspondente a β_j. A estatística T tem distribuição $t - Student$ com $n - k - 1$ graus de liberdade. Note-se que este teste é um teste marginal uma vez que o cálculo do coeficiente de regressão β_j é condicionado pela existência de todas as outras variáveis regressoras, $x_i (i \neq j)$, do modelo. Assim, este teste testa a contribuição relativa de x_j no modelo.

2.4 Medidas de qualidade do ajustamento do modelo

Depois de se confirmar que o modelo ajustado é significativo pode-se avaliar a qualidade do seu ajustamento.

Coeficiente de determinação

As medidas de associação, designadas normalmente por coeficientes de correlação, quantificam a intensidade e a direção da associação entre duas variáveis (correlação simples) ou entre uma variável dependente e um conjunto de outras variáveis (correlação múltipla). Genericamente, um coeficiente de correlação ρ varia entre os limites -1 e 1 e pode ser interpretado de acordo com a Tabela 2.1. Quando $\rho > 0$ diz-se que a relação é positiva, quando $\rho < 0$ diz-se negativa.

Tabela 2.1: Coeficiente de correlação.

$\lvert\rho\rvert$	Relação
0	nula
$]0; 0.20[$	muito fraca
$[0.20; 0.40[$	fraca
$[0.40; 0.70[$	moderada
$[0.70; 0.90[$	forte
$[0.90; 1[$	muito forte
1	perfeita

O coeficiente de correlação que quantifica a intensidade e a direção de associação do tipo linear entre duas variáveis quantitativas é o coeficiente de Pearson, R. Na regressão linear multipla utilizam-se, normalmente, duas medidas baseadas neste coeficiente: o coeficiente de determinação, R^2, e o coeficiente de determinação ajustado, R^2_a. Estas medidas são um caso particular do coeficiente de correlação de Pearson pois medem a correlação entre uma variável dependente e duas ou mais variáveis independentes. Têm a particularidade de variarem apenas entre 0 e 1, isto é, medem apenas a intensidade da relação, no entanto os seus valores podem ser interpretados da mesma forma que os coeficientes de correlação.
O R^2 é dado por (Murteira et al., 2010):

$$R^2 = \frac{VE}{VT}, \tag{2.22}$$

e mede a proporção de variabilidade total que é explicada pelo modelo de regressão, isto é, mede a qualidade de ajustamento do modelo aos dados da amostra. De um modo geral, R^2 aumenta quando um regressor é adicionado ao modelo, independentemente da sua influência sobre a variável resposta. Assim sendo, é difícil avaliar se o aumento do R^2 está a indicar algo

de relevante. Como tal, alguns investigadores preferem utilizar em alternativa o R_a^2 pois este só aumenta quando a variável regressora adicionada ao modelo reduz o QME relativamente à variância total. O coeficiente de determinação ajustado pode ser calculado por (Murteira et al., 2010)

$$R_a^2 = R^2 - \frac{k(1-R^2)}{n-k-1}. \tag{2.23}$$

O R_a^2 é ainda um melhor estimador do coeficiente de determinação da população do que R^2 e pode ser interpretado como uma medida de capacidade de generalização do modelo para outras amostras da mesma população.

Erro Quadrático Médio

Uma outra medida que pode ser utilizada para avaliar a qualidade de ajustamento do modelo é o erro quadrático médio, dado por

$$EQM = \frac{1}{n-k+1}\sum_{i=1}^{n}(y_i - \widehat{y}_i).$$

É usual utilizar a sua raiz quadrada $REQM = \sqrt{\frac{1}{n-k+1}\sum_{i=1}^{n}(y_i - \widehat{y}_i)}$, como medida do erro de previsão, pois tem as mesmas unidades da variável resposta, y.

2.5 Intervalos de Confiança (IC)

Depois de verificar que o modelo estimado é significativo e se ajusta bem aos dados, um procedimento útil para estimação, ou controlo, é a utilização de intervalos de confiança para os parâmetros, valor médio e previsões de futuras observações.

IC para os coeficientes de regressão

Para construir intervalos de confiança para os β_j supõe-se que os erros são independentes e normalmente distribuídos com valor médio nulo e variância constante, σ^2.

Da equação (2.21), um intervalo de confiança a $100(1-\alpha)\%$ para os coeficientes de regressão $\beta_j, j = 0, 1, ..., k$, é

$$\left[\widehat{\beta}_j - t_{\frac{\alpha}{2},n-k+1}\sqrt{\widehat{\sigma}^2 C_{jj}};\ \ \widehat{\beta}_j + t_{\frac{\alpha}{2},n-k+1}\sqrt{\widehat{\sigma}^2 C_{jj}}\right], \tag{2.24}$$

onde $t_{\frac{\alpha}{2},n-k+1}$ é o valor crítico da distribuição t-Student com $n-k+1$ graus de liberdade no percentil $\frac{\alpha}{2}$.

Estes intervalos são obtidos com base nas estatísticas definidas na subsecção relativa aos testes de hipóteses. Note-se que nessa subsecção apresenta-se o teste para a nulidade dos coeficientes de regressão individualmente, onde aceitar a hipótese nula é equivalente a verificar que zero pertence ao intervalo de confiança definido na equação (2.24) para cada coeficiente de regressão.

IC para o valor médio de y

Uma das aplicações mais utilizadas dos modelos de regressão é a estimação de um valor médio de y para um conjunto de observações particular das variáveis $x_1, x_2, ..., x_k$. Seja

$$\boldsymbol{x}_0^T = \begin{bmatrix} 1 & x_{01} & x_{02} & \dots & x_{0k} \end{bmatrix}^T$$

uma observação particular das variáveis $x_1, x_2, ..., x_k$. Pretende-se estimar

$$\theta = E[y_0|\boldsymbol{x}_0] = \beta_0 + \beta_1 x_{01} + \beta_2 x_{02} + ... + \beta_k x_{0k}. \tag{2.25}$$

por

$$\widehat{\theta} = \widehat{\beta}_0 + \widehat{\beta}_1 x_{01} + \widehat{\beta}_2 x_{02} + ... + \widehat{\beta}_k x_{0k}. \tag{2.26}$$

Prova-se que (Murteira et al., 2010)

$$Var[\widehat{\theta}|\boldsymbol{X}, \boldsymbol{x}_0] = Var[\widehat{\boldsymbol{\beta}}\boldsymbol{x}_0|\boldsymbol{X}, \boldsymbol{x}_0] = \sigma^2 \boldsymbol{x}_0^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_0. \tag{2.27}$$

A partir da estatística

$$\frac{\widehat{\theta} - \theta}{\sqrt{\widehat{\sigma}^2 \boldsymbol{x}_0^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_0}} \tag{2.28}$$

que segue uma distribuição t-Student com $n - k + 1$ graus de liberdade é possível construir um intervalo de confiança, a $100 \times (1 - \alpha)$, para a média de y no ponto $\boldsymbol{x}_0$ da seguinte forma

$$\left[\widehat{\theta} - t_{\frac{\alpha}{2}, n-k+1} \sqrt{\widehat{\sigma}^2 \boldsymbol{x}_0^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_0};\ \ \widehat{\theta} + t_{\frac{\alpha}{2}, n-k+1} \sqrt{\widehat{\sigma}^2 \boldsymbol{x}_0^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_0}\right]. \tag{2.29}$$

IC para a previsão de novas observações

Outra aplicação muito utilizada dos modelos de regressão é a previsão de futuras observações em y correspondentes a valores particulares das variáveis regressoras. Seja novamente

$$\boldsymbol{x}_0^T = \begin{bmatrix} 1 & x_{01} & x_{02} & \dots & x_{0k} \end{bmatrix}^T$$

uma observação particular das variáveis regressoras. Pode-se obter uma estimativa pontual de y substituindo os valores de $\boldsymbol{x}_0$ na equação do modelo ajustado:

$$\widehat{y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 x_{01} + \widehat{\beta}_2 x_{02} + ... + \widehat{\beta}_k x_{0k}. \tag{2.30}$$

Esta estimativa é afetada pelo erro aleatório, pelo que a variância do valo médio de y não serve para construir o intervalo de confiança para a previsão de novas observações.

Seja

$$e_0 = y_0 - \widehat{y}_0 \tag{2.31}$$

o erro de previsão. Prova-se que (Murteira et al., 2010)

$$Var[e_0|\boldsymbol{X}, \boldsymbol{x}_0] = Var[y_0 - \widehat{y}_0|\boldsymbol{X}, \boldsymbol{x}_0] = \sigma^2[1 + \boldsymbol{x}_0^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_0]. \tag{2.32}$$

A partir da estatística

$$\frac{\widehat{y}_0 - y_0}{\sqrt{\widehat{\sigma}^2(1 + \boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0)}}, \tag{2.33}$$

que segue uma distribuição t-Student com $n-k+1$ graus de liberdade, é possível construir um intervalo de previsão pontual a $100\times(1-\alpha)\%$ para uma observação futura de y correspondente a um observação particular das variáveis regressoras, $\boldsymbol{x}_0$, da seguinte forma

$$\left[\widehat{y}_0 - t_{\frac{\alpha}{2},n-k+1}\sqrt{\widehat{\sigma}^2(1 - \boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0)};\ \ \widehat{y}_0 + t_{\frac{\alpha}{2},n-k+1}\sqrt{\widehat{\sigma}^2(1 - \boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0)}\right]. \tag{2.34}$$

2.6 Validação dos pressupostos do modelo

O modelo de regressão linear só pode ser utilizado para estimar parâmetros e fazer inferência se os seus pressupostos forem validados.

A análise dos pressupostos permite concluir que estes não se verificam ou que não há evidências para supor que sejam violados. Esta última situação não significa que está a ser feita uma validação correta dos pressupostos no entanto não há razões para afirmar que está incorreta. Assim, mesmo que o modelo seja significativo e todos os seus pressupostos sejam verificados não significa que o modelo seja adequado mas tão só um modelo plausível. Se o modelo for declarado não adequado, torna-se necessário definir um modelo diferente.

Ao longo do texto foram feitas referências aos pressupostos do modelo de regressão: no cálculo das estimativas dos parâmetros é referido que os erros ϵ do modelo têm de ser não correlacionados, com valor médio nulo e variância constante (homocedasticidade) e que para fazer testes de hipóteses e construir intervalos de confiança se pressupõe ainda que os erros do modelo ϵ seguem uma distribuição normal. Mais, as variáveis regressoras devem ser não colineares e medidas sem erros.

De seguida são apresentadas algumas técnicas para verificar estes pressupostos.

Diagnóstico da independência dos erros

A independência dos erros pode ser verificada através do teste de hipóteses de Durbin-Watson. As hipóteses deste teste são:

$$H_0 : \rho_{e_{i+1},e_i} = 0 \qquad \textit{vs.} \qquad H_1 : \rho_{e_{i+1},e_i} \neq 0 \text{ para algum } i$$

onde ρ_{e_{i+1},e_i} representa a auto correlação entre e_{i+1} e e_i. A estatística de teste, sob H_0, é dada por

$$d = \frac{\sum_{i=1}^{n-1}(e_{i+1} - e_i)^2}{\sum_{i=1}^{n} e_j^2}.$$

Esta estatística toma valores no intervalo $[0, 4]$. Se $d \approx 2$ pode concluir-se que não há evidência de auto-correlação entre os erros (Marôco, 2010). Field (2009) sugere que, de um forma geral no modelo de regressão múltipla, estes valores não devem ser inferiores a um ou superiores a três.

Diagnóstico de normalidade dos erros

A normalidade dos erros aleatórios pode ser analisada através de:

i. Gráfico de probabilidade normal (PP-plot)

   Neste gráfico é possível visualizar a distribuição de probabilidades dos valores observados com os valores esperados segundo uma distribuição Normal (representados por uma linha em diagonal). Caso as observações registadas se aproximem dessa diagonal, sem nenhum afastamento significativo, como na Figura 2.1 (a), não há evidências para rejeitar que os erros aleatórios seguem uma distribuição Normal. Caso as observações se afastem significativamente da diagonal, como na Figura 2.1 (b), há evidência estatística para rejeitar a normalidade dos erros aleatórios.

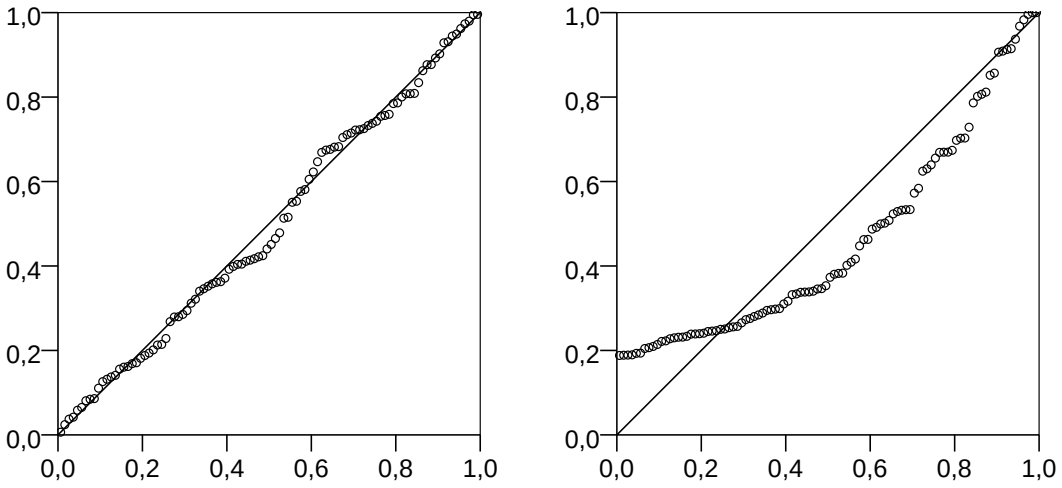
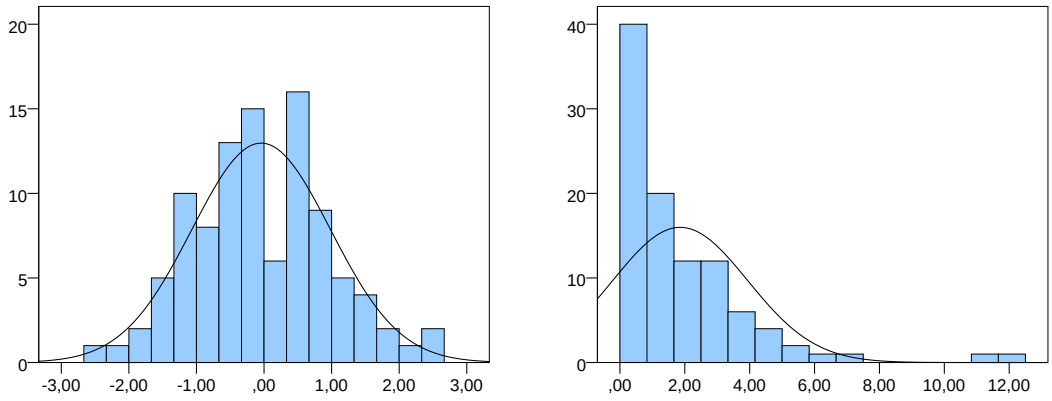


Figura 2.1: PP-plot dos erros da regressão: (a) com distribuição aproximadamente Normal e (b) com distribuição não Normal.

ii. Histograma dos erros *standardizados*

   Ao histograma dos erros *standardizados* sobrepõe-se a curva Normal e procuram-se afastamentos evidentes das barras do histograma à forma simétrica e unimodal da curva. Caso não existam diferenças notórias, não há evidências para rejeitar a normalidade dos erros aleatórios.



(a) (b)

Figura 2.2: Histograma de erros de regressão: (a) com distribuição aproximadamente Normal e (b) com distribuição não Normal.

iii. Teste de Kolmogorov-Smirnov

Para inferir analiticamente sobre a suposição da normalidade dos erros pode aplicar-se o teste de qualidade de ajustamento de Kolmogorov-Smirnov.

As hipóteses avaliadas são:

$$H_0 : \text{os dados seguem uma distribuição Normal}$$
$$\textit{vs.}$$
$$H_1 : \text{os dados não seguem uma distribuição Normal.}$$

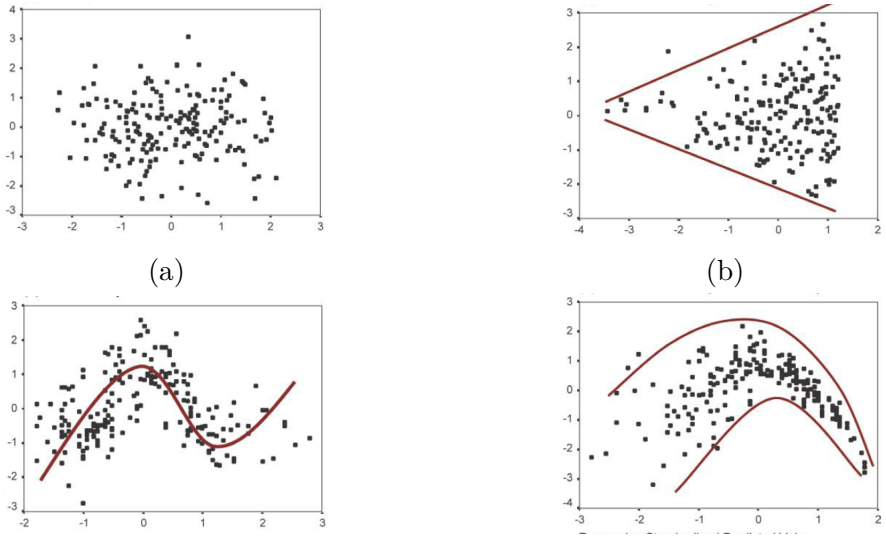
Este teste observa a diferença máxima em valor absoluto entre a função distribuição acumulada para os dados, neste caso a Normal, e a função de distribuição empírica dos dados, e compara essa diferença com um valor crítico, para um dado nível de significância.

Importa referir que na presença de grandes amostras este teste tende a rejeitar a hipótese nula com elevada frequência, como referem Hall, Neves e Pereira (2011).

Diagnóstico de homocedasticidade dos erros

Uma das técnicas utilizadas para verificar o pressuposto de que os erros são homocedásticos é a análise do gráfico dos resíduos *standardizados vs.* valores ajustados *standardizados*. Se a nuvem de pontos se apresenta limitada por uma banda aproximadamente horizontal, dispersa de forma aparentemente aleatória, como na Figura 2.3a, então não há evidências para colocar em causa a homocedasticidade dos erros.

Se a nuvem de pontos apresenta uma forma afunilada, como na Figura 2.3b, pode indicar a presença de heterocedasticidade nos erros. Se a forma da nuvem for curvilínea, como na Figura 2.3c, indica não-linearidade na relação entre a variável resposta e as variáveis preditoras.



(a) (b) (c) (d)

Figura 2.3: Gráficos exemplificativos dos erros de regressão *standardizados* (eixo y) *vs.* valores ajustados *standardizados* (eixo x). Em (a) não há evidências para por em causa a homocedasticidade dos erros de regressão, em (b), (c) e (d) os erros de regressão não são homocedásticos. Adaptado de Field (2009).

Se a nuvem apresenta ambas as formas, como na Figura 2.3d, pode indica a presença de heterocedasticidade e não-linearidade.

Este pressuposto pode ser verificado analiticamente através do teste de White simplificado (Murteira et al., 2010). As hipóteses deste teste são:

$$H_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_n^2 \qquad vs. \qquad H_1 : \sigma_i^2 \neq \sigma_j^2 \text{ para algum } (i,j) \text{ com } i \neq j.$$

Para efetuar este teste é necessário fazer a regressão de e^2 sobre $\widehat{y}$ e $\widehat{y}^2$, denominada de regressão auxiliar. A estatística de teste é dada por

$$W_s = nR_{aux}^2,$$

onde R_{aux}^2 é o coeficiente de determinação da regressão auxiliar. Esta estatística de teste segue uma distribuição χ^2 com dois graus de liberdade.

Diagnóstico de colinearidade e multicolinearidade

É importante garantir que as variáveis regressoras não estão correlacionadas entre si ou que apresentam correlações fracas. Se existirem correlações fortes entre duas ou mais variáveis regresssoras, diz-se que existe multicolinearidade. Na presença de multicolinearidade exata não é possível estimar o modelo de regressão, pois a equação (2.12) não tem solução. Quando duas ou mais variáveis são fortemente (mas não exatamente) correlacionadas, é possível estimar o modelo de regressão, no entanto a resolução da equação (2.12) torna-se difícil devido à

ocorrência de problemas de instabilidade numérica, conduzindo a estimadores de elevada variância, o que torna um modelo desprovido de valor.

O diagnóstico de multicolinearidade pode ser efetuado recorrendo à análise das correlações entre pares de variáveis independentes. Correlações elevadas ($R > |0,75|$) conduzem normalmente a problemas de multicolinearidade, no entanto, quando mais do que dois regressores estão envolvidos numa relação de dependência quase-linear, esta análise não garante que não existam correlações elevadas nessas relações. Uma forma de contornar este problema é recorrer aos valores de *Variance Inflation Factor* (VIF) e *tolerance*. Field (2009) sugere que os valores de VIF não devem ser superiores a 10 e que os valores de *tolerance* não devem ser inferiores a $0,2$ para cada variável no modelo, e que a média dos valores de VIF seja aproximadamente 1(um), de forma garantir que a multicolinearidade não causará problemas no cálculo das estimativas dos coeficientes de regressão.

2.7 Incorporação de variáveis artificiais no modelo de regressão linear

As variáveis utilizadas no modelo são, de um modo geral, quantitativas. No entanto existem alguns fatores que se pretendem inserir no modelo que não podem ser representados por uma variável quantitativa. Se alguma variável independente é nominal é possível inclui-la no modelo através de variáveis artificiais ou *dummy*.

Para modelar um fator com m modalidades, define-se uma modalidade como *acontecimento de referência* e constroem-se $m-1$ variáveis binárias, para as restantes, evitando desta forma a existência de colinearidade exata entre estas variáveis que serão regressores. Os coeficientes de regressão associados às variáveis artificiais são interpretados como a diferença em relação à alternativa escolhida para referência. No modelo de regressão devem constar todas as modalidades do fator.

O fator qualitativo pode ter efeitos na constante ou num qualquer regressor quantitativo. Além do mais, quando são introduzidos dois ou mais fatores no modelo devem ter-se em conta as interações entre eles. Estes temas não serão abordados neste relatório, no entanto, podem ser consultados em Montgomery, Peck e Vining (2006).

Nos modelos de regressão utilizados mais adiante neste relatório os efeitos das variáveis artificiais serão considerados apenas no termo independente e não serão analisadas interações entre os regressores.

2.8 Seleção de regressores

Na maioria das situações práticas não se consegue especificar à partida qual o conjunto ideal de regressores que deve ser incluído na equação de regressão de forma a obter o melhor modelo.

Existindo um conjunto de variáveis potencialmente úteis para explicar o comportamento da variável resposta, pretende-se selecionar aquelas que são efetivamente relevantes.

Método exaustivo

Este procedimento requer que o investigador ajuste um modelo de regressão a todas as possíveis combinações de regressores disponíveis e escolher, de acordo com algum critério, o que produz melhores resultados (por exemplo, maior R^2 ou menor EQM). Assumindo que a constante é incluída em todos os modelos testados e se existem k variáveis candidatas a regressoras, será necessário ajustar 2^k modelos de regressão.

Fazer o ajustamento a todas as combinações possíveis de variáveis pode ser muito pesado computacionalmente. No sentido de agilizar este procedimento, foram criados métodos que avaliam os modelos de cada vez que se adicionam ou removem variáveis, constituindo uma aproximação ao método exaustivo. Esses métodos são apresentados em seguida.

Seleção *forward*

No método de seleção *forward*, o modelo é iniciado apenas com a constante, sem regressores, e ao qual se vai inserindo uma variável regressora de cada vez. A primeira variável regressora a ser selecionada para entrar no modelo é a que apresenta uma maior correlação com a variável resposta. A segunda a ser selecionada é aquela que tem uma maior correlação parcial com a variável resposta, ou seja, a correlação obtida excluindo o efeito da variável já incorporada no modelo. O processo continua até que não existam mais variáveis que possam ser incluídas no modelo cuja correlação parcial com a variável resposta seja estatisticamente significativa ou até que todos os potenciais regressores sejam incluídos no modelo.

Seleção *backward*

No método de seleção *backward* o modelo é iniciado com todas as variáveis regressoras disponíveis e, a cada passo, um regressor cuja presença no modelo não contribua para explicar uma porção significativa da variação total da variável resposta é removido. O método termina quando não se justifique remover mais variáveis regressoras do modelo ou quando não existam regressores na equação.

Seleção *stepwise*

O método de seleção *stepwise* é um híbrido dos dois métodos de seleção anteriores. Inicia com a seleção *forward* mas a adição de uma variável é avaliada pelo método *backward*, pelo que se um regressor adicionado num passo anterior se torna redundante é eliminado do modelo e se um regressor que foi eliminado num passo anterior se torna relevante, é inserido novamente no modelo. O modelo final contém apenas variáveis preditoras consideradas relevantes para

explicar a variável resposta.

Nos métodos de seleção *forward* e *backward*, a inclusão ou exclusão de regressores tem um carácter definitivo. Um regressor que seja incluído no modelo utilizando a seleção *forward* nunca mais o abandona, mesmo que mais tarde se torne supérfluo como consequência da entrada de novos regressores; um regressor que seja excluído do modelo utilizando a seleção *backward* não volta a ser selecionado, mesmo que mais tarde se torne útil, consequência da eliminação de outros regressores do modelo. Contrariamente ao que acontece nestes dois métodos de seleção de variáveis, na seleção *stepwise* a cada passo os regressores já introduzidos no modelo são reexaminados de forma a garantir que o modelo contém apenas regressores considerados relevantes pelo método. Esta característica levou à escolha deste último método para selecionar as variáveis regressoras a considerar nos modelos de regressão construídos mais adiante neste relatório.

2.9 Validação do modelo

Embora o ajustamento do modelo de regressão seja importante, não deve ser o critério único de avaliação do modelo. É importante que as estimativas dos coeficientes de regressão sejam representativas dos efeitos parciais das variáveis regressoras na população. Como tal, é necessário verificar se as estimativas dos coeficientes de regressão traduzem uma relação entre as variáveis regressoras e a variável resposta que faça sentido de acordo com a teoria subjacente ao problema. Coeficientes com sinais diferentes daqueles que se esperam em teoria podem indiciar que o modelo é inapropriado (Snee, 1997).
Uma outra forma de avaliar o modelo é verificar se este faz boas previsões. Para isso é usual utilizarem-se técnicas de validação cruzada (ou *cross-validation*), existindo para tal várias metodologias. A que será utilizada mais adiante neste relatório é a proposta por Tabachnick e Fidell (2007) e consiste no seguinte: a amostra é dividida em duas partes, onde cerca de 80% das observações, o chamado conjunto de treino, são utilizadas para obter o modelo de regressão e as restantes 20%, o chamado conjunto de teste, são utilizadas para validar o modelo; o modelo obtido com o conjunto de treino é utilizado para prever os valores da variável resposta no conjunto de teste. São obtidos e comparados os coeficientes de determinação para ambos os conjuntos: uma grande discrepância entre os valores de R^2_{treino} e R^2_{teste} indica um sobre-ajustamento do modelo ao conjunto de treino e os resultados da análise do modelo de regressão não devem ser generalizados. Esta análise é um pouco subjetiva já que não estão definidos valores a partir dos quais se possa considerar que a diferença entre os coeficientes de determinação dos conjuntos de treino e teste seja considerada elevada colocando em causa o ajuste do modelo.

2.10 Análise de observações influentes

Um *outlier* é uma observação atípica, que tem um comportamento diferente da maioria das observações e, por isso, apresenta erros *standardizados* superiores, em valor absoluto, aos erros das demais observações. De forma a identificar essas observações é comum *standardizar* os erros fazendo

$$e'_i = \frac{e_i}{\widehat{\sigma}^2},$$

de forma a que a média se mantenha zero e o desvio-padrão seja unitário. Assim, se os erros forem normalmente distribuídos, aproximadamente 95% dos erros *standardizados* são menores do que 2 e aproximadamente 99% são menores do que 3, em valor absoluto. As observações com erros *standardizados* maiores do que 3 em valor absoluto devem ser examinadas com atenção pois indicam valores atípicos.

Os efeitos dos *outliers* podem ser moderados ou extremos, consoante se encontram no meio do domínio das observações ou nos limites do domínio das observações, respetivamente. Nem todos os *outliers* afetam os parâmetros e estatísticas da reta de regressão, por isso, para além de detetar a existência de *outliers* é importante detetar os pontos que afetam de forma significativa o ajuste do modelo.

Uma medida de disgnóstico, sugerida por Cook em 1977, que mede a influência da observação $\boldsymbol{x}_i$ sobre a estimação de $\boldsymbol{\beta}$ é a distância de Cook (Montgomery, Peck e Vining, 2006):

$$DC_i = \frac{(\widehat{\boldsymbol{\beta}}_{-i} - \widehat{\boldsymbol{\beta}})^T \boldsymbol{X}^T \boldsymbol{X} (\widehat{\boldsymbol{\beta}}_{-i} - \widehat{\boldsymbol{\beta}})}{(k+1) EQM},$$

onde $\widehat{\boldsymbol{\beta}}_{-i}$ é a estimativa do vetor de parâmetros $\boldsymbol{\beta}$ sem considerar a observação i.

Cook considera que observações com DC_i superiores a 1 são excessivamente influentes.

É importante para o investigador avaliar a existência de observações influentes para conhecer a sua amostra. Estes casos, apesar de serem diferentes da maioria, podem apresentar valores plausíveis e não se justificar a sua remoção da amostra. No entanto, se a amostra apresentar muitos casos influentes é aconselhável utilizar técnicas de regressão robusta, pois não são tão afetadas pela existência deste tipo de observações como o método dos mínimos quadrados.

Capítulo 3

Problema proposto

O problema proposto pela empresa consiste em obter um modelo de previsão dos preços de venda para imóveis situados na cidade de Lisboa, com base nas características dos mesmos. A empresa pretende ainda identificar os atributos que efetivamente acrescentam valor ao imóvel.

Neste capítulo são apresentados os passos efetuados na resolução do problema proposto e as conclusões obtidas. Primeiramente são apresentados os procedimentos efetuados para a limpeza e preparação da base de dados, seguidos de uma análise descritiva para compreender, descrever e resumir a base de dados. Finalmente, são apresentados os modelos de regressão aplicados, a sua avaliação, interpretação e conclusões.

3.1 Tratamento da base de dados

A base de dados cedida pela empresa foi extraída da plataforma *Imovirtual* no dia 17 de fevereiro de 2015 e é constituída por 7853 imóveis com o código-postal de Lisboa.

A informação sobre cada imóvel contempla um número para identificação (ID), o tipo de negócio (venda, arrendamento permanente ou arrendamento para férias), o código-postal, a morada, uma breve descrição, o preço, a condição (novo, renovado, usado, em construção, para recuperar ou ruína), a natureza (apartamento, moradia, prédio, quarto ou terreno), a tipologia, a área útil, a área bruta, o número de casas de banho, o ano de construção, o certificado energético e um campo onde pode ser assinalado se o imóvel tem as seguintes características: acessibilidade a pessoas com mobilidade condicionada, alarme, aquecimento central, ar condicionado, árvores de fruto, condomínio fechado, cozinha equipada, elevador, estacionamento, garagem, hidromassagem/jacuzzi, jardim, lareira, mobilado, piscina, quintal/horta, imóvel do banco, som ambiente, varanda, vigilância/segurança, vista de campo/serra, vista de mar e vista de rio.

Como o objetivo do trabalho é modelar o preço de venda de imóveis para habitação, foram selecionados apenas os apartamentos e moradias que se encontravam indicados para venda, perfazendo um total de 5930 imóveis. De seguida foram eliminadas as observações com ID duplicado e com valores para venda sem sentido, o que reduziu a amostra para 3980 imóveis.

Para além das características estruturais do imóvel que constam na base de dados cedida pela empresa, considerou-se necessário criar variáveis que permitissem avaliar a sua localização. Para tal foram consideradas as categorias de pontos de interesse constantes da plataforma *Livin'X*: paragens de autocarro, estações de metro, cafés, mercearias/supermercados, ginásios, hospitais, *pub's*, parques, escolas, farmácias e restaurantes; e consideraram-se ainda os aterros sanitários e zonas industriais do distrito de Lisboa e o complexo químico de Algés. De forma a incluir esta informação no modelo de regressão foram determinadas as distâncias mínimas de automóvel, e tempos correspondentes, de cada um dos imóveis ao ponto mais próximo de cada categoria. Estas distâncias e tempos foram estimados tomando por base o código-postal do imóvel uma vez que a sua localização exata não estava disponível. Foi possível obter as distâncias e respetivos tempos aos pontos acima descritos para 1471 imóveis.

Para melhor caracterizar a localização dos imóveis criaram-se duas variáveis: o número de pontos de interesse num raio de um quilómetro e o número de bairros sociais por freguesia. A primeira soma as categorias de pontos de interesse da plataforma Livin'X que se encontram num raio de um quilómetro em torno do imóvel, a segunda foi obtida através da carta BIP/ZIP da cidade de Lisboa de 2013.

Limitações das variáveis recolhidas

A informação contida na base de dados foi extraída de uma plataforma *online* de acesso gratuito, onde se podem publicar quaisquer imóveis com a informação que o vendedor considere pertinente, sem que seja necessária a confirmação da informação disponibilizada. Neste sentido, a informação disponibilizada apresenta algumas limitações que não podem ser contornadas e que a seguir se listam:

1. Ausência de exclusividade na publicação do anúncio de venda de um imóvel e não identificação do seu proprietário. Os imóveis podem ser inseridos pelo proprietário e/ou por agências imobiliárias, pelo que o mesmo imóvel pode surgir repetidas vezes publicado na plataforma sem haver uma forma de o identificar;

2. A publicação de um imóvel tem um objetivo comercial, o que pode levar o seu publicitador a omitir ou alterar atributos que possam desvalorizar o imóvel;

3. A existência de campos abertos e campos não obrigatórios permitem a adição de informação potencialmente relevante, no entanto, dada a sua natureza, os imóveis não podem ser avaliados da mesma forma;

4. A existência de campos abertos pode conduzir ainda a incoerências entre estes e os campos fechados;

5. O preço dos imóveis considerado não é o valor real de transação, podendo este estar sub ou sobre estimado pelo publicitador;

6. A localização exata do imóvel não está acessível, pelo que o cálculo das distâncias e tempos aos tipos de pontos de interesse é aproximado;

7. Os pontos de interesse utilizados foram apenas os disponíveis na plataforma *Livin'X*.

8. As distâncias e tempos apenas informam sobre a proximidade à categoria de ponto de interesse, não sendo possível identificar quantos pontos de interesse do mesmo tipo ficam na proximidade. Por exemplo, se a distância ao café é menor do que um quilómetro isto indica que existe pelo menos um café num raio de um quilómetro mas não dá informação sobre quantos cafés existem próximos do imóvel.

9. Falta de garantia da representatividade da amostra.

Uma forma de evitar incoerências e a falta de informação na inserção de um imóvel na plataforma *online* podia passar por definir os campos de preenchimento como obrigatórios e fechados, sempre que possível, e que a localização do imóvel fosse um dos dados a preencher ainda que pudesse ficar apenas na base de dados da plataforma e não visível para os utilizadores visitantes. Outra solução para contornar alguns dos problemas acima enumerados podia passar por fazer uma parceria com uma agência imobiliária. Desta forma era possível evitar que o mesmo imóvel fosse considerado na base de dados em duplicado, pois a agência tem conhecimento da sua localização; os imóveis seriam avaliados da mesma forma, segundo as normas da agência; não existiriam incoerências de informação já que os imóveis são observados *in loco*; era possível ter conhecimento do verdadeiro valor de transação dos imóveis após a sua venda.

A recolha de dados é um aspeto fundamental na análise de regressão. Não basta recolher muitas observações com um elevado número de variáveis, é necessário recolher dados relevantes e fiáveis.

3.2 Análise Descritiva

Nesta secção é apresentada uma análise descritiva da base de dados com o objetivo de caraterizar a amostra, apresentando a informação das suas variáveis sumariada, para que rapidamente sejam identificadas as caraterísticas da sua distribuição, relações e padrões, com recurso a gráficos e estatísticas.

Será efetuada uma análise dos dados em bruto, começando por apresentar uma breve análise univariada para as variáveis disponíveis, seguida de uma análise bivariada de pares de variáveis cuja análise foi considerada pertinente.

Os atributos dos imóveis foram divididos em três grupos: **Atributos gerais** composto pelas variáveis apresentadas na Tabela 3.1, **Atributos descritivos** constituído pelas variáveis binárias da Tabela 3.2, **Atributos de localização** constituído pelas variáveis da Tabela 3.3.

Tabela 3.1: Variáveis que descrevem os atributos gerais do imóvel.

Variável	Descrição
Preço	preço em euros
ÁreaÚtil	em m^2
ÁreaBruta	em m^2
Idade	em anos, dada pela diferença entre 2015 e o ano de construção
Natureza	apartamento ou moradia
Estado	novo ou usado
NrQuartos	número de quartos
NrWCs	número de casas de banho
CE	certificado energético (A+, A, B, B-, C, D, E, F)
NrCaracterísticas	número de características da tabela 3.2

Tabela 3.2: Variáveis binárias que identificam os atributos descritivos do imóvel.

Acessibilidades*	Alarme	AquecimentoCentral
ArCondicionado	ÁrvoreFruto	CondomínioFechado
CozinhaEquipada	Elevador	Estacionamento
Garagem	Hidromassaagem	Jardim
Lareira	Mobilado	Piscina
Quintal	ImóvelBanco	SomAmbiente
Varanda	Vigilância	VistaCampoSerra
VistaMar	VistaRio	

*: acessibilidade a pessoas com mobilidade condicionada

Tabela 3.3: Variáveis que descrevem os atributos de localização dos imóveis.

Variável	Descrição
NrBairrosSociais	nº de bairros sociais
NrPontosInteresse	nº de tipo de pontos de interesse no raio de 1Km
TempoAutocarro	tempo à estação de autocarro
TempoMetro	tempo à estação de metro
TempoCafé	tempo ao café
TempoMercearia	tempo à mercearia ou supermercado
TempoGinásio	tempo ao ginásio
TempoHospital	tempo ao hospital
TempoPub	tempo ao clube de diversão noturna
TempoParque	tempo ao parque
TempoEscola	tempo à escola
TempoFarmácia	tempo à farmácia
TempoRestaurante	tempo ao restaurante
TempoAterro	tempo ao aterro
TempoZI	tempo à zona industrial
TempoCQ	tempo ao complexo químico de Algés

Análise Univariada

A amostra é constituída por 1471 imóveis dos quais 87,4% são apartamentos e os restantes são moradias; 35,1 % dos imóveis são novos, 58,9 % são usados e os restantes não têm indicação sobre o seu estado (valores omissos).

As características das variáveis quantitativas contínuas relativas aos atributos gerais em estudo são apresentadas na Tabela 3.4.

Tabela 3.4: Estatísticas descritivas das variáveis quantitativas contínuas da amostra pertencentes ao grupo dos atributos gerais do imóvel.

		Preço	ÁreaÚtil	ÁreaBruta	Idade	NrQuartos	NrWCs
N	Validos	1471	1471	613	800	1471	1111
	Omissos	0	0	858	671	0	360
Mediana		485000,00	160,00	195,00	1,00	3,00	2,00
Média		824723,43	214,22	323,00	14,04	3,57	2,45
Desvio-padrão		834980,38	192,69	406,69	22,62	1,97	1,32
Assimetria		2,07	3,88	4,72	1,84	1,15	,97
Curtose		8,04	27,57	36,20	2,56	1,91	1,22
Amplitude		7760500	2282	4981	120	10	7
Mínimo		39500	18	19	0	0	1
Máximo		7800000	2300	5000	120	10	8

A variável *Preço* distribui-se num intervalo de valores aceitável, no entanto o seu desvio-padrão é da ordem do 800 mil euros, o que indica uma grande dispersão relativamente ao valor médio. Este é também muito elevado pois está influenciado pela existência de muitos *outliers* à direita como se pode ver na Figura 3.1a. Existem imóveis com preços muito baixos, menores que 93000 euros, que são na sua maioria apartamentos usados, com muitos anos e pequenas áreas e existem também imóveis com preços muito elevados, maiores que 2300000 euros, que são maioritariamente apartamentos novos com áreas generosas e certificados energéticos de classes A e superior ou moradias com muitos quartos e amplas áreas. Estes imóveis, que representam cerca de 50% da amostra, apesar de terem características algo diferentes dos restantes, apresentam preços demasiado elevados. Estas observações podem corresponder a imóveis com características muito particulares que não foram analisadas, como palácios, palacetes ou outras características únicas que podem estar na origem de preços tão elevados. Podem ser ainda imóveis que não foram avaliados realisticamente. A mediana da variável preço é aproximadamente 485 mil euros, ou seja, quase metade dos imóveis da amostra são de gama alta [1].

As áreas útil e bruta distribuem-se num intervalo de valores aceitável, no entanto a amplitude desses intervalos é elevada. A amostra contém imóveis com áreas muito pequenas que são na sua maioria apartamentos T0 ou T1 e áreas muito grandes que são na sua maioria

[1]Uma forma de obter vistos Golden Visa é fazer um investimento imobiliário de valor igual ou superior a 500 mil euros, pelo que foi admitido que a partir deste valor os imóveis são de gama alta.

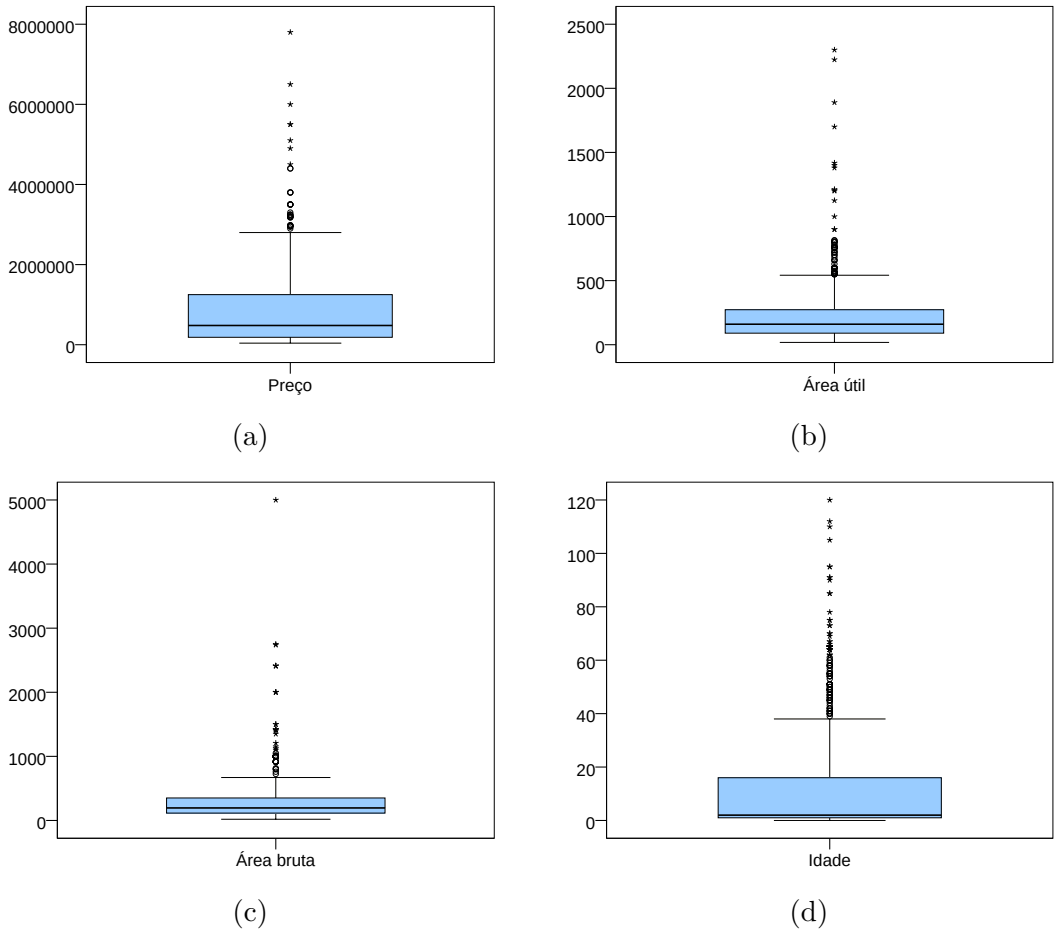


Figura 3.1: Distribuição do (a) Preço, (b) área útil, (c) área bruta e (d) idade dos imóveis.

moradias com muitos quartos e com preços na ordem dos milhões de euros. Estas variáveis apresentam uma distribuição assimétrica positiva e muitos *outliers* como se pode ver nas Figuras 3.1b e 3.1c.

A maioria dos imóveis são recentes, tendo um ano ou menos, existindo no entanto imóveis muito antigos: cerca de 10% do imóveis tem mais de 65 anos e alguns ultrapassam os 100 anos (ver Figura 3.1d).

A **ÁreaBruta** tem 858 valores omissos, cerca de 58% da amostra, e a **Idade** tem 671 valores omissos, cerca de 46% da amostra. Estes campos são de preenchimento não obrigatório na plataforma, o que pode estar na causa da existência desta elevada percentagem de valores omissos.

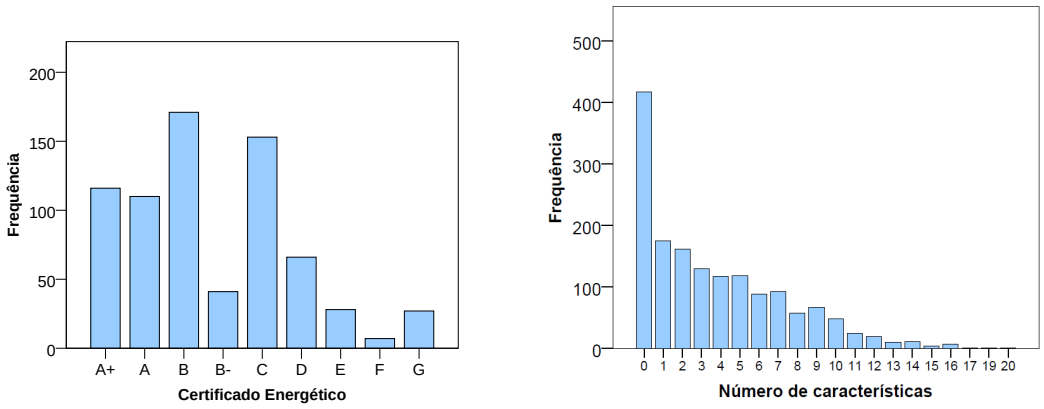
A variável **NrQuartos** toma valores entre 0 e 10, que é um intervalo de valores aceitável para esta variável; os imóveis sem quartos são apartamentos e apenas duas moradias têm um quarto. No entanto existem alguns apartamentos com mais de oito quartos, valores acima do expectável.

O **NrWCs** varia entre 1 e 8, que é um intervalo aceitável, já que todos os imóveis têm

pelo menos uma casa de banho. No entanto, existe uma pequena percentagem de imóveis com mais do que seis casas de banho, valores muito acima do expectável. Esta variável apresenta 360 valores omissos, cerca de 25% da amostra.

Observa-se na Figura 3.2a que mais de 50 % dos imóveis estão avaliados nas categorias de certificação energética B ou superiores, que pode ser resultado de uma crescente preocupação com os gastos energéticos, não só a nível ambiental mas também de custos, já que a maior parte dos imóveis novos têm certificado A ou A+. Esta variável apresenta 779 valores omissos representando cerca de 53% da amostra. É também um campo de preenchimento não obrigatório na plataforma, o que pode ser a causa deste tão elevado número de valores omissos.

O gráfico de barras com as frequências relativas da variável **NrCaracterísticas** (ver Figura 3.2b) mostra que existem muitos imóveis em que não é referido nenhum atributo descritivo e que nenhum imóvel combina todos eles. O campo da plataforma Imovirtual onde se selecionam os atributos descritivos do imóvel é de preenchimento não obrigatório e, como tal, esses atributos não estarem assinalados não significa que o imóvel não os tenha efetivamente. Os atributos descritivos mais observados são **CozinhaEquipada**, **Elevador**, **Estacionamento** e **Varanda**, e os menos observados são **ÁrvoresFruto**, **VistaMar**, **Mobilado** e **Quintal** (ver Figura 3.3).



(a) (b)

Figura 3.2: Frequência das variáveis (a)**CE** e (b)**NrCaracterísticas**.

As variáveis relativas aos tempos de viagem às categorias de pontos de interesse consideradas distribuem-se em intervalos de valores plausíveis com amplitudes razoáveis e não apresentam valores omissos. Da análise da Tabela A.1 do Apêndice A conclui-se que a maior parte dos imóveis estão próximos dos pontos de interesse estação de autocarro, café, mercearia, ginásio, hospital, parque, *pub* e estação de metro, encontrando-se a menos de três minutos; e que a maior parte dos imóveis estão afastados do complexo químico, dos aterros sanitários e das zonas industriais, mais de oito minutos. Esta relação é expectável, pois os imóveis estão localizados na cidade de Lisboa, uma cidade com rápidos acessos a estes serviços. Como seria

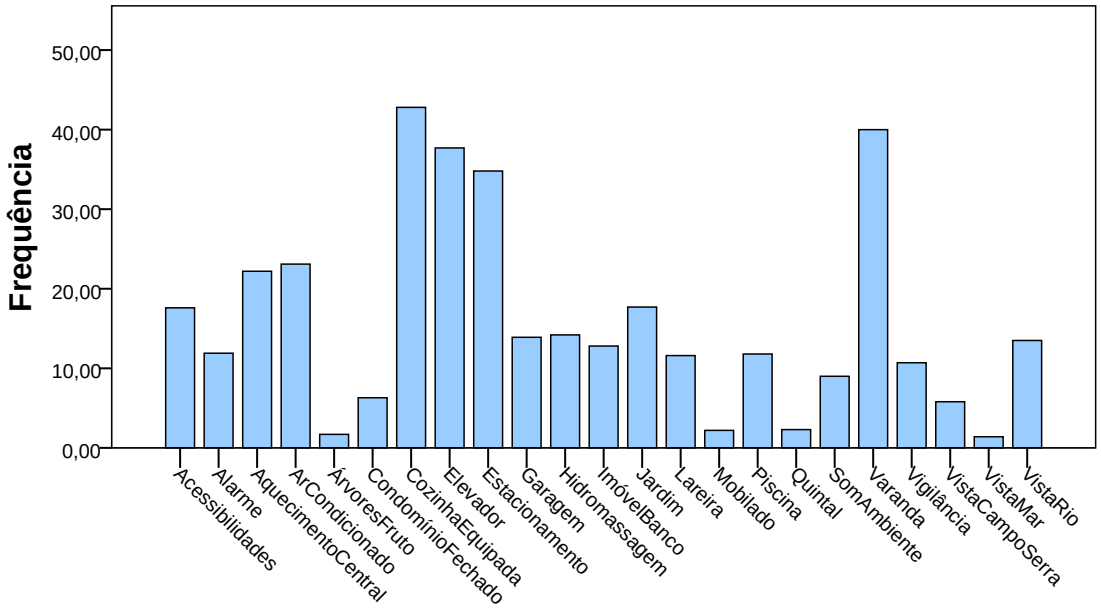


Figura 3.3: Frequência dos atributos descritivos do imóvel.

de esperar, os imóveis estão mais afastados do complexo químico, aterros sanitários e zonas industriais, do que dos restantes pontos de interesse.

As distâncias de viagem aos pontos de interesse não serão analisadas já que apresentam uma correlação forte com os tempos de viagem e apresentam muitos valores nulos, pelo que se considerou que os tempos fossem mais informativos. No entanto as distâncias foram utilizadas para construir a variável **NrPontosInteresse**.

Análise Bivariada

A análise bivariada tem como objetivo descrever e sumariar a informação sobre pares de variáveis, recorrendo a métodos gráficos e coeficientes de correlação.

As Tabelas 3.5 e 3.6 apresentam os coeficientes de correlação de Pearson entre as variáveis quantitativas da amostra. A Figura 3.4 apresenta os gráficos de dispersão entre alguns pares de variáveis.

Podem retirar-se as seguintes conclusões:

- O **Preço** tem uma relação linear positiva muito forte com a **ÁreaÚtil**, verificando-se um aumento do preço do imóvel com o aumento da sua área útil, um comportamento bastante expectável.

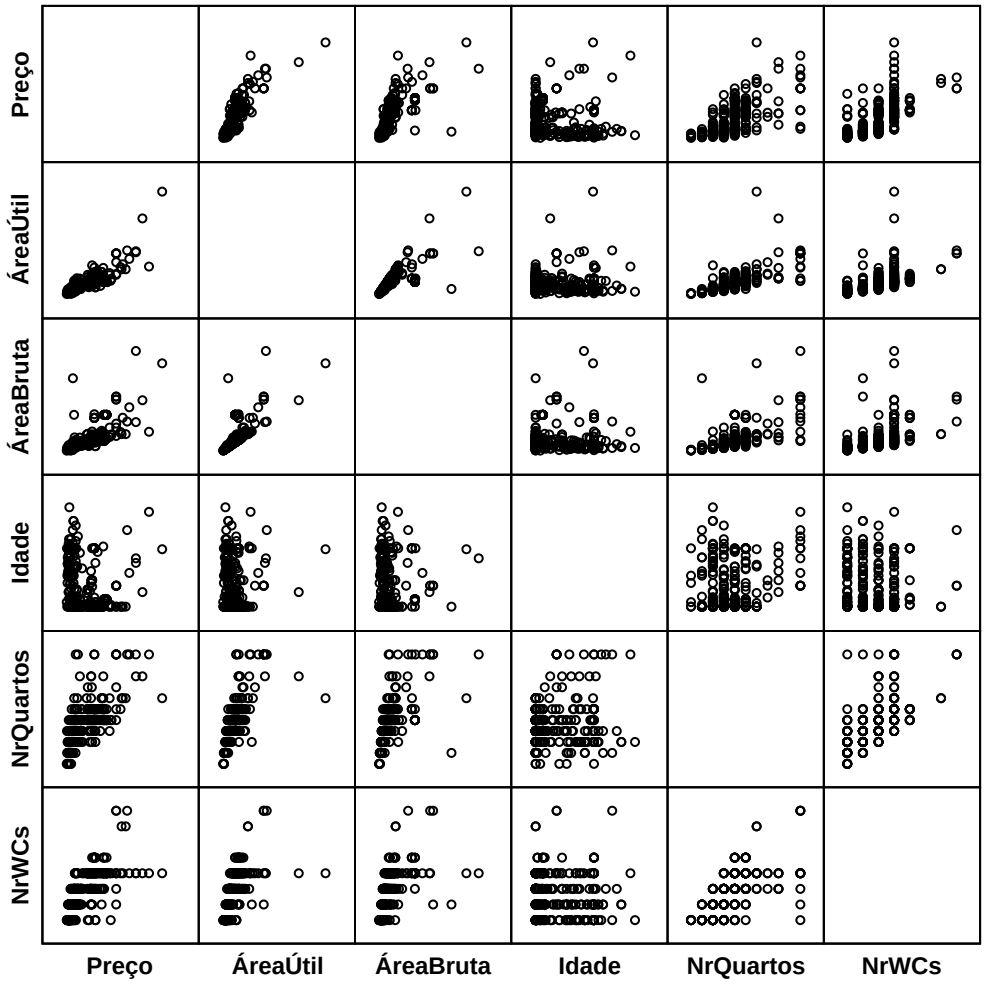


Figura 3.4: Gráficos de dispersão para **Preço**, **ÁreaÚtil**, **ÁreaBruta**, **Idade**, **NrQuartos** e **NrWCs**.

- O preço do imóvel diminui com a idade, embora o coeficiente de correlação não seja muito elevado. A relação entre estas variáveis é expectável, pois o preço do imóvel diminui com o aumento da idade, no entanto a relação linear é fraca, o que pode indicar que a idade de um imóvel não é um fator relevante para a formação do seu preço, já que um imóvel pode ser antigo mas estar renovado.

- A variável **Preço** aumenta com o número de quartos e o número de casas de banho.

- As correlações mais elevadas são observadas entre as variáveis **ÁreaÚtil** e **ÁreaBruta**; entre estas variáveis existe uma relação linear positiva muito forte o que é um resultado bastante previsível, já que para alguns imóveis o valor da área bruta coincide com o valor da área útil.

- Existe uma relação linear positiva forte entre a **ÁreaÚtil** e as variáveis **NrQuartos** e **NrWCs**.

- A variável **Preço** e os tempos de viagem a categorias de pontos de interesse apresentam uma relação linear fraca. Analisando os coeficientes significativos, observa-se que o afastamento temporal aos cafés, mercearias, ginásios, hospitais, restaurantes e escolas diminui o preço dos imóveis, como esperado. Constata-se que a proximidade aos aterros sanitários e zonas industriais diminui o preço dos imóveis, uma relação também expectável. Contrariamente ao esperado, o **TempoComplexoQuímico** apresenta uma relação linear negativa com o **Preço**, i.e. o preço dos imóveis aumenta com a proximidade ao complexo químico.

Tabela 3.5: Coeficientes de correlação de Pearson entre as variáveis **Preço**, **ÁreaÚtil**, **ÁreaBruta**, **Idade**, **NrQuartos** e **NrWCs**.

	Preço	ÁreaÚtil	ÁreaBruta	Idade	NrQuartos	NrWCs
Preço	1					
ÁreaÚtil	0,807**	1				
ÁreaBruta	0,733**	0,813**	1			
Idade	-0,148**	0,012	0,05	1		
NrQuartos	0,671**	0,706**	0,611**	0,133**	1	
NrWCs	0,729**	0,673**	0,556**	-0,14**	0,698**	1

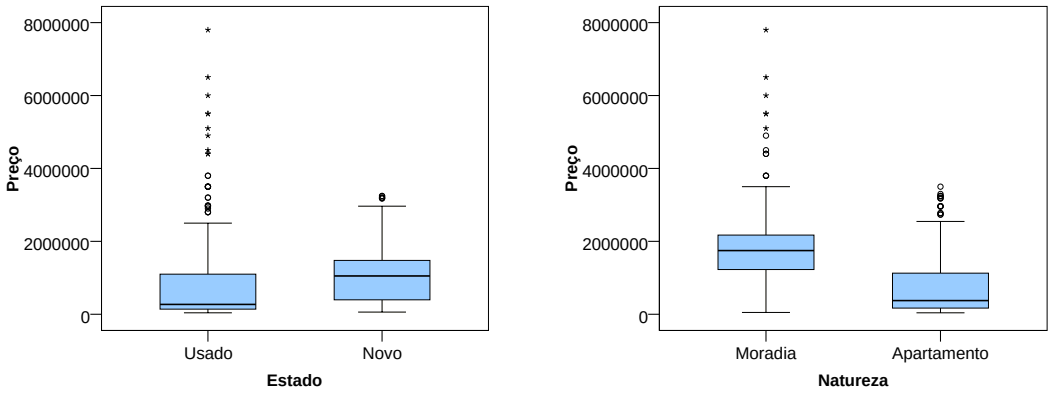
**: a correlação é significativa ao nível de 0.01

Tabela 3.6: Coeficientes de correlação de Pearson entre a variável **Preço** e as variáveis relativas aos tempos.

	Preço
TempoAutocarro	0,044
TempoCafé	-0,053*
TempoMercearia	-0,060*
TempoGinásio	-0,158**
TempoHospital	-0,119**
TempoPub	0,006
TempoParque	0,010
TempoFarmácia	0,036
TempoComplexoQuímico	-0,082**
TempoRestaurante	-0,110**
TempoEscola	-0,122**
TempoMetro	0,158**
TempoAterro	0,138**
TempoZonaIdustrial	0,020

*: a correlação é significativa ao nível de 0.05
**: a correlação é significativa ao nível de 0.01

Verifica-se que a maior parte das moradias apresenta preços superiores à maioria dos apartamentos. O que sugere que sugere que estes possam ter um comportamento diferente ao

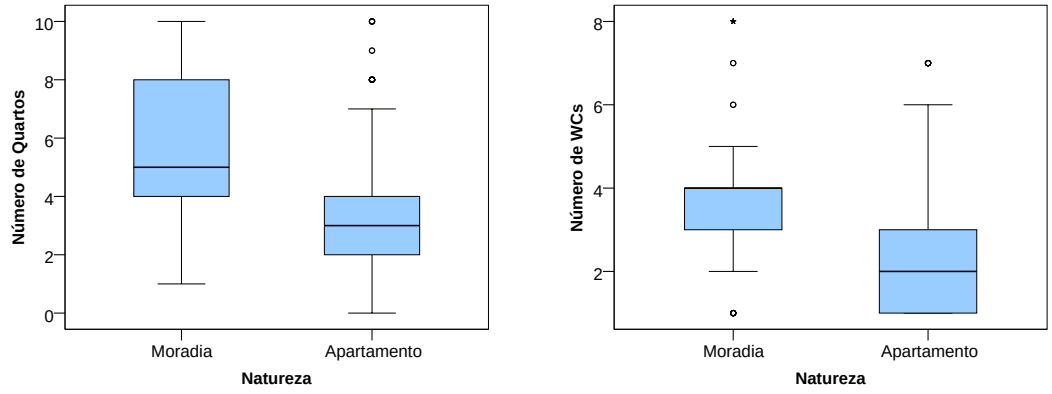


(a) (b)

Figura 3.5: Distribuição do Preço em função do (a) Estado e da (b) Natureza.

Tabela 3.7: Tabela de contingência das variáveis Natureza e Estado.

		Natureza		
		Moradia	Apartamento	Total
Estado	Usado	153	714	867
	Novo	14	502	516
	Total	167	1216	1383



(a) (b)

Figura 3.6: Distribuição do **NrQuartos** e **NrWCs** em função da natureza do imóvel.

nível do preço. Estas diferenças não são tão notórias quando são comparados os imóveis novos e usados (ver Figura 3.5). Os gráficos da Figura 3.6 reforçam a ideia de que as características das moradias e apartamentos têm comportamentos diferentes.

Segundo Ribeiro (*Preço das casas de luxo em Lisboa aumenta 54% influenciado pelos Golden Visa*) nos últimos anos os preços dos imóveis para habitação da gama alta em Lisboa têm sido inflacionados pelas compras para efeitos de obtenção de Golden Visa (mecanismo que dá direito a residir em Portugal a quem faz investimentos imobiliários a partir de 500 mil

euros), o que sugere que a avaliação destes imóveis seja diferente relativamente aos de gama inferior.

3.3 Análise de Regressão

Nesta secção são apresentados os resultados da aplicação do modelo de regressão linear múltipla, sem interações entre os regressores nem restrições. Os modelos de regressão foram obtidos utilizando o método de seleção de variáveis *stepwise* e escolhida a opção *listwise* que apenas utiliza as observações sem valores omissos nas suas variáveis. Considerou-se ainda a probabilidade de erro tipo I de 5% em todas as análises.

Os modelos apresentados pretendem avaliar a relação do preço dos imóveis com os seus atributos gerais, descritivos e de localização, descritos na secção anterior. Dos atributos acessíveis, a **ÁreaBruta** e a **Idade** não serão incluídos na construção dos modelos já que a elevada percentagem de valores omissos destas variáveis iria reduzir bastante o tamanho da amostra; além disso, a **ÁreaBruta** tem uma correlação forte com a **ÁreaÚtil** e a **Idade** uma correlação fraca com o **Preço**.

Com base na análise descritiva dos dados, o problema foi abordado de cinco formas diferentes a seguir enlencadas, considerando:

1. todos os imóveis num só modelo - GERAL;
2. apenas os apartamentos - APARTAMENTOS;
3. apenas as moradias - MORADIAS;
4. os imóveis com preço inferior a 500 mil euros - GAMA MÉDIA/BAIXA ;
5. os imóveis com preço superior ou igual a 500 mil euros - GAMA ALTA.

De forma a encontrar o conjunto de atributos que melhor descreve o problema, em cada uma das abordagens foram testadas três combinações de variáveis independentes, que podem ser consultadas na Tabela 3.8. Em todos os modelos são incluídos os atributos gerais e de localização pois são aqueles que a empresa pretende compreender. Após uma análise preliminar dos modelos, entendeu-se que os atributos descritivos são importantes para explicar a variação do preço dos imóveis, pelo que se procurou a forma mais adequada de os incluir no modelo de regressão.

O grupo dos atributos descritivos é composto por 23 variáveis dicotómicas que, avaliadas separadamente, podem explicar pouco o comportamento do preço dos imóveis, pois a informação que dão ao modelo está muito detalhada. Nesse sentido, considerou-se o agrupamento destas variáveis de forma a dar um maior contributo para o modelo, considerando características mais abrangentes e não particularizar demasiado (Neto, 2008). Assim, estas variáveis foram agrupadas de duas formas distintas:

1. Uma só variável, resultado da soma do número de atributos descritivos de cada imóvel (**NrCaracterísticas**)

2. Cinco novas variáveis que agrupam os atributos descritivos dos imóveis da seguinte forma: **Segurança** (soma dos atributos Alarme, CondomínioFechado e Vigilância), **Vistas** (soma das vistas de rio, mar e campo/serra), **Outside** (soma dos atributos ÁrvoresFruto, Piscina, Varanda, Jardim e Quintal), **Inside** (soma dos atributos Acessibilidade, Aquecimento, Hidromassagem, Lareira, ArCondicionado, Elevador, Mobilado e SomAmbiente) e **Estacionamento** (soma dos lugares de garagem e estacionamento).

De forma a construir um modelo para cada abordagem foi aplicada a técnica de validação cruzada, cujo procedimento pode ser consultado no Apêndice B. Após a aplicação da validação cruzada foi selecionado um modelo para cada abordagem cujas estatísticas sumárias se encontram na Tabela 3.9. Curiosamente o modelo que foi construído com todas as variáveis tem maiores valores de R^2 e $R^2_{ajustado}$ do que os modelos das outras abordagens que são menos abrangentes. Apesar disso, os restantes modelos têm resultados melhores para os imóveis específicos da sua abordagem. Esta conclusão é fundamentada nos resultados apresentados na Tabela 3.10 que resultam da aplicação do modelo GERAL aos conjuntos de imóveis dos modelos: APARTAMENTOS, MORADIAS, GAMA MÉDIA/BAIXA e GAMA ALTA. Verifica-se efetivamente que os modelos criados em específico para cada um destes tipos de imóveis produzem melhores resultados do que o modelo GERAL (comparar Tabelas 3.9 e 3.10).

De seguida serão analisados cada um dos modelos em particular.

Tabela 3.8: Combinações de variáveis independentes.

		C1	C2	C3
Atributos Gerais	ÁreaÚtil	×	×	×
	Estado	×	×	×
	Natureza	×	×	×
	NrQuartos	×	×	×
	NrWCs	×	×	×
Atributos Descritivos	NrCaracterísticas		×	
	Acessibilidades	×		
	Alarme	×		
	AquecimentoCentral	×		
	ArCondicionado	×		
	ÁrvoresFruro	×		
	CondomínioFechado	×		
	CozinhaEquipada	×		
	Elevador	×		
	Estacionamento	×		
	Garagem	×		
	Hidromassagem	×		
	ImóvelBanco	×		×
	Jardim	×		
	Lareira	×		
	Mobilado	×		
	Piscina	×		
	Quintal	×		
	SomAmbiente	×		
	Varanda	×		
	VistaMar	×		
	Vigilancia	×		
	VistaCampoSerra	×		
	VistaRio	×		
	Vistas			×
	Segurança			×
	Outside			×
	GaragemLugar			×
	Inside			×
Atributos de Localização	NrBairrosSociais	×	×	×
	NrPontosInteresse	×	×	×
	TempoAutucarro	×	×	×
	TempoCafé	×	×	×
	TempoMercearia	×	×	×
	TempoGinásio	×	×	×
	TempoHospital	×	×	×
	TempoPub	×	×	×
	TempoParque	×	×	×
	TempoFarmácia	×	×	×
	TempoCQ	×	×	×
	TempoRestaurante	×	×	×
	TempoEscola	×	×	×
	TempoMetro	×	×	×
	TempoAterro	×	×	×
	TempoZI	×	×	×

Tabela 3.9: Estatísticas sumárias dos modelos de regressão linear múltipla para cada abordagem.

Modelo	R^2	$R^2_{ajustado}$	REQM
GERAL	0,817	0,815	310354,5
APARTAMENTOS	0,808	0,806	238034,7
MORADIAS	0,667	0,654	594309,5
GAMA MÉDIA/BAIXA	0,746	0,738	59939,1
GAMA ALTA	0,655	0,648	424607,3

Tabela 3.10: Coeficiente de determinação do modelo GERAL aplicado aos imóveis de cada uma das abordagens.

Imóveis	R^2
APARTAMENTOS	0,796
MORADIAS	0,656
GAMA MÉDIA/BAIXA	0,579
GAMA ALTA	0,645

Modelo GERAL

Dos modelos construídos a partir de todos os imóveis da amostra, aquele que produziu melhores resultados foi o construido a partir do conjunto de variáveis independentes C1. O método *stepwise* selecionou como preditores significativos do comportamento médio da variável **Preço** as variáveis apresentadas na Tabela 3.11.

Tabela 3.11: Tabela sumária do modelo GERAL.

	β_j	Erro padrão	β'_j	t	p
(Constante)	136205,1	54110,7		2,517	0,012
ÁreaÚtil	2698,5	83,9	0,617	32,166	<0,001
Estado	166020,4	22173,9	0,108	7,487	<0,001
Natureza	-220589,9	39413,0	-0,100	-5,597	<0,001
NrWCs	96593,5	10686,3	0,173	9,039	<0,001
Piscina	330268,4	40763,8	0,129	8,102	<0,001
Elevador	-55669,5	21489,7	-0,038	-2,591	0,010
Jardim	117363,8	30581,4	0,063	3,838	<0,001
Vigilância	62234,7	31472,0	0,028	1,977	0,048
VistaRio	101381,4	27356,9	0,051	3,706	<0,001
TempoPub	-22529,1	4819,0	-0,064	-4,675	<0,001
NrBairrosSociais	-13232,7	5767,175	-0,031	-2,295	0,022

Avaliando os valores absolutos dos coeficientes de regressão *standardizados*, verifica-se que os preditores que apresentam maior contribuição relativa para a explicação do comportamento médio do preço de venda dos imóveis são a **ÁreaÚtil**, o **NrWCs** e a **Piscina**, seguindo-se o **Estado** e a **Natureza**. Os preditores com menor impacto são **TempoPub**, **Jardim**, **VistaRio**, **Elevador**, **NrBairrosSociais** e **Vigilância**.

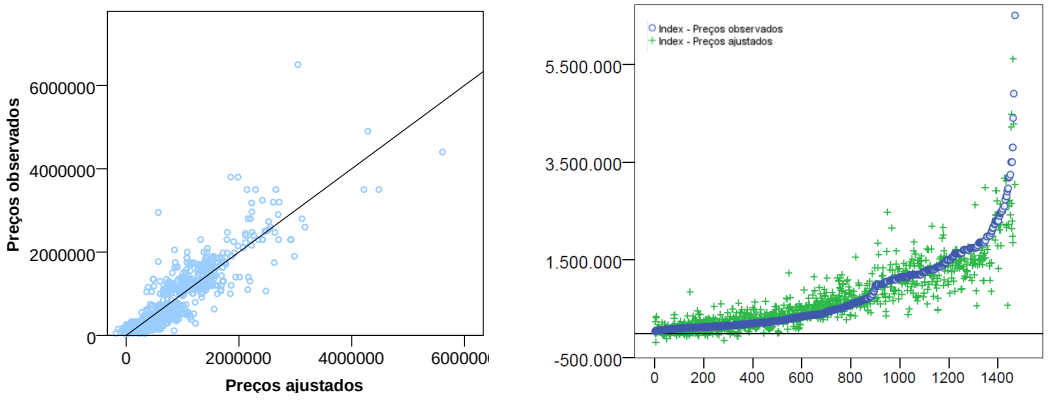
O modelo final ajustado fica então:

$$\begin{aligned}\widehat{Preço} =& 136205,1 + 2698,5\boldsymbol{ÁreaÚtil} + 166020,4\boldsymbol{Estado} - 220589,9\boldsymbol{Natureza} \\ &+ 96593,5\boldsymbol{NrWCs} + 330268,4\boldsymbol{Piscina} - 55669,5\boldsymbol{Elevador} \\ &+ 117363,8\boldsymbol{Jardim} + 62234,7\boldsymbol{Vigilância} + 101381,4\boldsymbol{VistaRio} \\ &- 22529,1\boldsymbol{TempoPub} - 13232,7\boldsymbol{NrBairrosSociais}. \end{aligned} \tag{3.1}$$

As variáveis selecionadas por este modelo explicam cerca de 81,5% da variabilidade do preço dos imóveis, em torno da sua média ($F \approx 429,382; R_a^2 \approx 0,815, p < 0,001$). A REQM para este modelo é igual a 310354,5.

A **ÁreaÚtil** apresenta um coeficiente de 2698,5 o que indica que aumentando um metro quadrado na área útil do imóvel o seu preço aumenta em média 2698,5 euros. Um imóvel com mais uma casa de banho é em média 96593,5 euros mais caro e a existência de **Piscina** faz com que o preço médio de um imóvel aumente 330268,4 euros. O **Estado** apresenta um coeficiente de 166020,4 o que indica que um imóvel novo com as mesmas características de um imóvel usado é em média 166020,4 euros mais caro. Um apartamento com as mesmas características de uma moradia é em média 220589,9 euros mais barato. Pode concluir-se ainda que a existência de elevador e a proximidade a bairros sociais diminui o valor do imóvel; a proximidade a locais de diversão noturna, a existência de jardim, vigilância e vista de rio aumenta o valor do imóvel.

Na Figura 3.7a verifica-se que os pontos no gráfico de dispersão dos preços observados e preços ajustados pelo modelo se distribuem em torno da reta $y = x$ com ligeiros afastamentos que se tornam mais notórios com o aumento dos preços, indicando que este modelo faz previsões piores para imóveis com preços mais elevados. Na Figura 3.7b comparam-se as curvas dos preços observados e dos preços ajustados pelo modelo. Apesar de um comportamento semelhante são visíveis afastamentos significativos.



(a) (b)

Figura 3.7: Gráficos dos preços observados e preços ajustados pelo modelo GERAL.

Verificação dos pressupostos

Analisaram-se os pressupostos do modelo de regressão linear múltipla, nomeadamente o da independência e normalidade dos erros. O primeiro foi validado com a estatística de teste de Durbin-Watson ($d \approx 1,312$). O pressuposto da normalidade dos erros foi analisado graficamente através das Figuras 3.8a e 3.8b, não tendo sido validado, pois os valores obtidos distam consideravelmente dos valores de referência. No entanto, Murteira et al. (2010) referem que na presença de grandes amostras o facto do pressuposto da normalidade dos erros não ser validado não anula a inferência sobre o modelo de regressão desde que os erros tenham média nula e sejam homocedásticos. Como por definição os erros têm valor médio nulo, basta verificar se estes são homocedásticos. A fim de validar esse pressuposto foi construído o gráfico da Figura 3.8c de onde se conclui que os erros não têm variância constante, conclusão que é reforçada pelo resultado do teste de White-simplificado ($W_s \approx 100,580 > \chi(0.95,2) \approx 5,991$). Contudo, para Tabachnick e Fidell (2007) o facto de os erros não serem homocedásticos não invalida a análise de regressão, apenas a enfraquece. Recorde-se que o estimador dos mínimos quadrados, nesta circunstância, continua a ser consistente e não enviesado, simplesmente deixa de ser o mais eficiente entre os estimadores lineares não enviesados.

Da análise da Tabela 3.12, verifica-se que não existem problemas de colinearidade entre as variáveis envolvidas no modelo, dado que os valores de VIF são inferiores a 10, sendo a sua média $1,463$, e os valores de *tolerance* superiores a $0,2$.

Tabela 3.12: Valores de *tolerance* e VIF das variáveis selecionadas pelo modelo GERAL.

Variáveis	Tolerance	VIF
ÁreaÚtil	0,470	2,127
Estado	0,834	1,199
Natureza	0,532	1,878
NrWCs	0,472	2,121
Piscina	0,676	1,479
Elevador	0,785	1,273
Jardim	0,651	1,536
Vigilância	0,851	1,174
VistaRio	0,887	1,127
TempoPub	0,922	1,085
NrBairrosSociais	0,917	1,091

Foram ainda analisados os *outliers* do modelo não tendo sido observados valores muito discrepantes dos valores teóricos. Também não foram observadas observações com distância de Cook superior a 1, pelo que nenhuma observação foi considerada excessivamente influente.

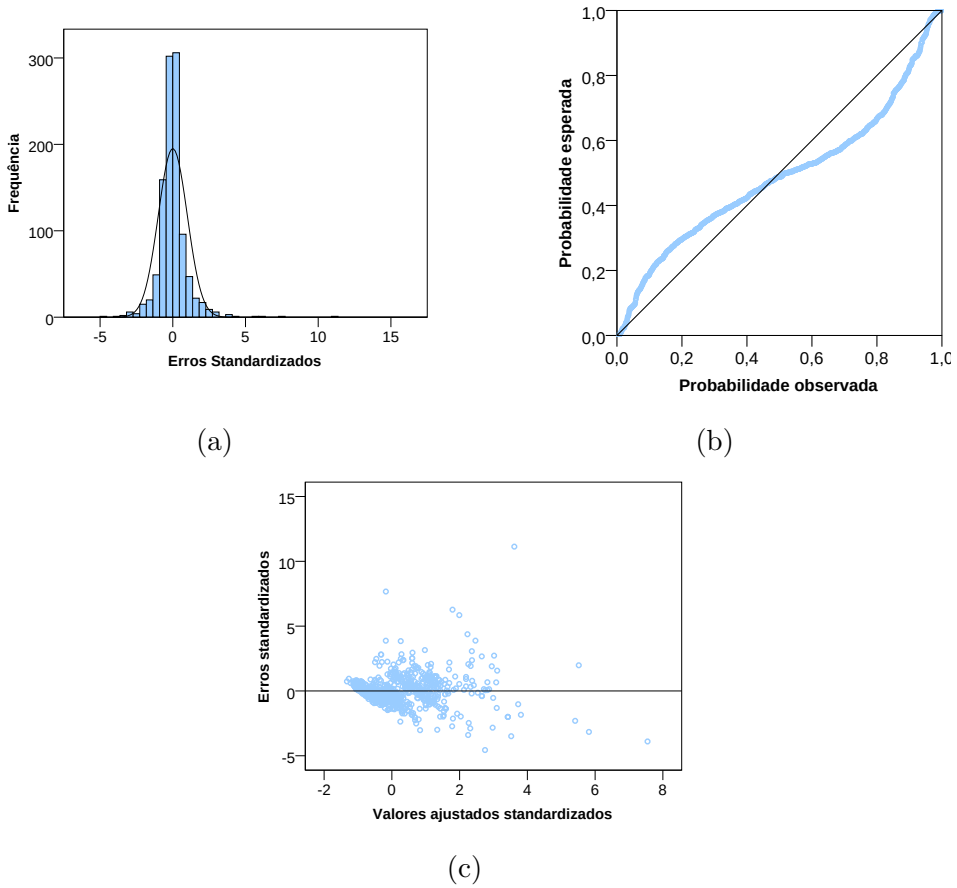


Figura 3.8: Gráficos para verificação de pressupostos do modelo dos erros do modelo GERAL: (a) histograma dos erros de regressão *standardizados*; (b) PP-plot dos erros de regressão *standardizados* e (c) verificação de homocedasticidade.

Modelo APARTAMENTOS

Dos modelos construídos a partir dos apartamentos da amostra, aquele que produziu melhores resultados foi o construído a partir do conjunto de variáveis independentes C1. As variáveis indicadas como significativas pelo modelo são apresentadas na Tabela 3.13.

Tabela 3.13: Tabela sumária do modelo APARTAMENTOS.

	$\widehat{\beta}_j$	Erro padrão	$\widehat{\beta}'_j$	t	p
(Constante)	-274201,6	44231,1		-6,199	<0,001
ÁreaÚtil	3427,1	112,9	0,695	30,359	<0,001
Estado	184889,3	17517,7	0,164	10,554	<0,001
Piscina	191591,8	40278,1	0,078	4,757	<0,001
NrWCs	49740,0	11043,5	0,106	4,504	<0,001
TempoAterro	8523,8	3132,0	0,041	2,721	0,007
Vigilância	87549,1	24714,7	0,054	3,542	<0,001
Jardim	77432,0	27069,9	0,047	2,860	0,004
NrBairrosSociais	-13160,9	4698,0	-0,042	-2,801	0,005
VistaRio	23104,1	21588,3	0,016	1,070	0,284

Verifica-se que as variáveis **ÁreaÚtil**, **Estado** e **NrWCs** apresentam maior contribuição

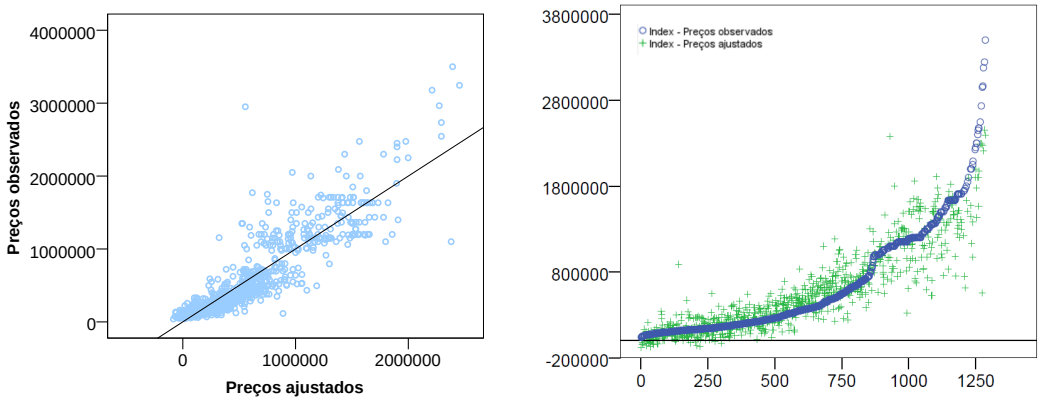
relativa para a explicação do comportamento médio do preço de venda dos apartamentos, seguindo-se a existência de **Piscina**, **Vigilância** e **Jardim**. Com um impacto menor tem-se os **NrBairrosSociais**, **TempoAterro** e **VistaRio**.

O modelo final ajustado fica então:

$$\begin{aligned}\widehat{Preço} = &-274201,6 + 3427,1\boldsymbol{ÁreaÚtil} + 184889,3\boldsymbol{Estado} + 191591,8\boldsymbol{Piscina} \\ &+ 49740,0\boldsymbol{NrWCs} + 8523,8\boldsymbol{TempoAterro} + 87549,1\boldsymbol{Vigilância} \\ &+ 77432,0\boldsymbol{Jardim} - 13160,9\boldsymbol{NrBairrosSociais} + 23104,1\boldsymbol{VistaRio}. \quad (3.2)\end{aligned}$$

As variáveis selecionadas por este modelo explicam cerca de 80,6% da variabilidade do preço dos apartamentos, em torno da sua média ($F \approx 433,979, R_a^2 \approx 0,806, p < 0,001$). A REQM para este modelo é igual a $238034,7$.

Na Figura 3.9a verifica-se que os pontos no gráfico de dispersão dos preços observados e preços ajustados pelo modelo se distribuem em torno da reta $y = x$, contudo apresentam afastamentos consideráveis à reta de referência. Da Figura 3.9b, verificamos que o comportamento dos preços ajustados segue a tendência dos observados, embora com maior variabilidade. Contudo são visíveis ligeiros afastamentos que indicam que o modelo não faz boas previsões do preço dos imóveis.



(a) (b)

Figura 3.9: Gráficos dos preços observados e ajustados pelo modelo APARTAMENTOS.

Pode ser consultado nos Apêndices C e D alguns dados que complementam a análise deste modelo, de entre os quais estatísticas e representações gráficas que servem para verificar os pressupostos do modelo. A partir da análise desses dados foi possível verificar que o pressuposto da independência dos erros é satisfeito, no entanto não foi validado o pressuposto da normalidade dos erros. Como acontece com o modelo anterior, o estimador dos mínimos quadrados continua a ser consistente e não enviesado, no entanto deixa de ser o mais eficiente entre os estimadores lineares não enviesados. Verificou-se ainda que não existem problemas de multicolinearidade entre as variáveis envolvidas no modelo.

Foram analisados os *outliers* do modelo não tendo sido observados valores muito discrepantes dos valores teóricos. Também não foram observadas observações com distância de Cook superior a 1, pelo que nenhuma observação foi considerada excessivamente influente.

Modelo MORADIAS

Dos modelos construídos para as moradias, aquele que produziu melhores resultados foi o construído a partir do conjunto de variáveis independentes C2. Foram selecionados cinco preditores significativos através do método de seleção *stepwise* que são apresentados na Tabela 3.14.

Tabela 3.14: Tabela sumária do modelo MORADIAS.

	$\widehat{\beta}_j$	Erro padrão	$\widehat{\beta}'_j$	t	p
(Constante)	67352,5	184802,9		0,365	0,716
ÁreaÚtil	2316,8	231,7	0,653	9,999	<0,001
TempoMetro	-54109,7	15389,0	-0,191	-3,516	<0,001
NrWCs	123657,0	40644,3	0,169	3,042	0,003
NrCaracterísticas	43765,8	12564,7	0,183	3,483	0,001
NrQuartos	35489,0	28213,8	0,086	1,258	0,211

Os preditores com maior contribuição relativa para a explicação do comportamento médio do preço de venda de moradias são a **ÁreaÚtil**, o **TempoMetro**, o **NrCaracterísticas**, o **NrWCs** e o **NrQuartos**.

O modelo final ajustado fica:

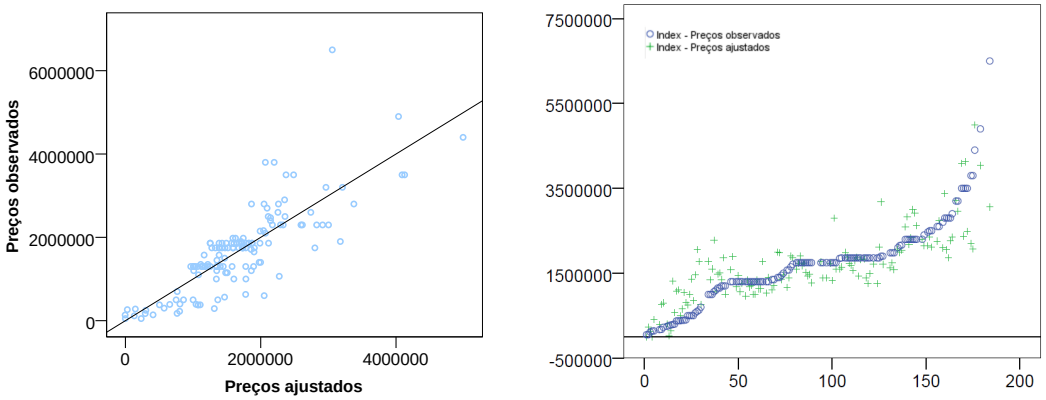
$$\begin{aligned}\widehat{Preço} =& 67352,5 + 2316,8\boldsymbol{ÁreaÚtil} - 54109,7\boldsymbol{TempoMetro} + 123657,0\boldsymbol{NrWCs} \\ &+ 43765,8\boldsymbol{NrCaracterísticas} + 35489,0\boldsymbol{NrQuartos}. \end{aligned} \tag{3.3}$$

As variáveis selecionadas por este modelo explicam cerca de 65,4% da variabilidade do preço dos apartamentos, em torno da sua média ($F \approx 52,783, R^2_a \approx 0,654, p < 0,001$). A REQM para este modelo é igual a $594309,5$.

O pressuposto da normalidade dos erros não foi verificado, contudo verifica-se a partir do PP-plot (ver Figura C.1 do Apêndice C) que têm um comportamento mais próximo da distribuição normal do que os erros dos modelos anteriores.

Nas Figuras 3.10a e 3.10b verifica-se que os pontos no gráfico de dispersão dos preços observados e preços preditos pelo modelo se distribuem em torno da reta $y = x$ com ligeiros afastamentos e que têm um comportamento semelhante. Comparativamente aos modelos anteriores, este faz melhores previsões.

Foram analisados os *outliers* do modelo não tendo sido observados valores muito discrepantes dos valores teóricos. Também não foram observadas observações com distância de Cook superior a 1, pelo que nenhuma observação foi considerada excessivamente influente.



(a) (b)

Figura 3.10: Gráficos dos preços observados e preços preditos pelo modelo MORADIAS.

Modelo GAMA MÉDIA/BAIXA

Dos modelos construídos para os imóveis de gama média/baixa, aquele que produziu melhores resultados foi o construído partir do conjunto de variáveis independentes C1. O método *stepwise* selecionou como preditores significativos do comportamento médio da variável **Preço** as variáveis apresentadas na Tabela 3.15.

Tabela 3.15: Tabela sumária do modelo GAMA MÉDIA/BAIXA.

	$\widehat{\beta}_j$	Erro padrão	$\widehat{\beta}'_j$	t	p
(Constante)	-12001,3	17603,4		-0,682	0,496
ÁreaÚtil	1194,2	100,1	0,441	11,933	<0,001
Estado	59324,5	6852,4	0,208	8,657	<0,001
AquecimentoCentral	29222,8	7521,3	0,092	3,885	<0,001
Elevador	19017,8	6145,3	0,080	3,095	0,002
NrBairrosSociais	-7148,4	1607,9	-0,114	-4,446	<0,001
TempoPub	-4024,1	1470,3	-0,069	-2,737	0,006
ArCondicionado	30679,8	7043,2	0,104	4,356	<0,001
NrQuartos	14682,3	3229,9	0,149	4,546	<0,001
CondomínioFechado	40242,9	14180,8	0,065	2,838	0,005
Estacionamento	29539,9	6785,2	0,113	4,354	<0,001
Quintal	-35984,6	14912,2	-0,050	-2,413	0,016
Garagem	26971,4	9774,0	0,061	2,760	0,006
NrWCs	17592,3	5037,1	0,109	3,493	0,001
TempoAterro	3897,7	1050,9	0,095	3,709	<0,001
TempoHospital	-5578,3	1820,9	-0,088	-3,063	0,002
TempoAutocarro	5543,3	2034,7	0,067	2,724	0,007
CozinhaEquipada	15326,9	5632,5	0,065	2,721	0,007
Piscina	41117,9	21188,3	0,041	1,940	0,053

Avaliando os valores absolutos dos coeficientes de regressão *standardizados* verifica-se que os preditores que apresentam maior contribuição relativa para a explicação do comportamento médio do preço de venda dos apartamentos são a **ÁreaÚtil**, o **Estado** e o **NrQuartos**, seguindo-se **NrBairrosSociais**, **Estacionamento**, **NrWCs** e **ArCondicionado**. Com menor contribuição relativa tem-se **TempoAterro**, **AquecimentoCentral**, **TempoHospital**, **Elevador**, **TempoPub**, **TempoAutocarro**, **CondomínioFechado**, **CozinhaEquipada**,

Garagem, **Quintal** e **Piscina**, por esta ordem de importância.

O modelo final ajustado é:

$$\begin{aligned}\widehat{Preço} = &-12001,3 + 1194,2\boldsymbol{ÁreaÚtil} + 59324,5\boldsymbol{Estado}\\ &+ 29222,8\boldsymbol{AquecimentoCentral} + 19017,8\boldsymbol{Elevador}\\ &- 7148,4\boldsymbol{NrBairrosSociais} - 4024,1\boldsymbol{TempoPub}\\ &+ 30679,8\boldsymbol{ArCondicionado} + 14682,3\boldsymbol{NrQuartos}\\ &+ 40242,9\boldsymbol{CondomínioFechado} + 29539,9\boldsymbol{Estacionamento}\\ &- 35984,6\boldsymbol{Quintal} + 26971,4\boldsymbol{Garagem}\\ &+ 17592,3\boldsymbol{NrWCs} + 3897,7\boldsymbol{TempoAterro}\\ &- 5578,3\boldsymbol{TempoHospital} + 5543,3\boldsymbol{TempoAutocarro}\\ &+ 15326,9\boldsymbol{CozinhaEquipada} + 41117,9\boldsymbol{Piscina}. \end{aligned} \tag{3.4}$$

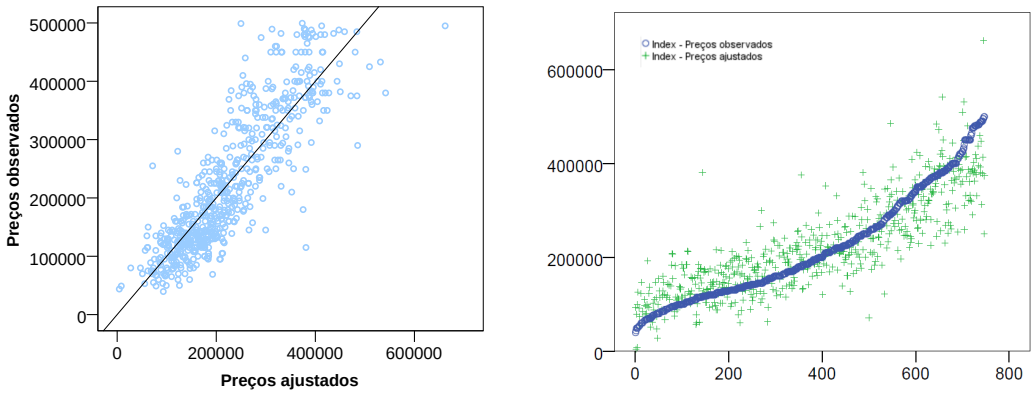
As variáveis selecionadas por este modelo explicam cerca de 73.8% da variabilidade do preço dos apartamentos, em torno da sua média ($F \approx 98,700, R_a^2 \approx 0,738, p < 0,001$). A REQM para este modelo é igual a $59939,1$.

Foi validado o pressuposto da independência dos erros de regressão, mas não o da normalidade (veja-se o Apêndice C). Verifica-se ainda que não há problemas de multicolinearidade entre as variáveis envolvidas no modelo.

Da Figura 3.11a, verificamos que o comportamento dos preços ajustados segue a tendência dos observados, embora com maior variabilidade.

Foram analisados os *outliers* do modelo não tendo sido observados valores muito discrepantes dos valores teóricos. Também não foram observadas observações com distância de Cook superior a 1, pelo que nenhuma observação foi considerada excessivamente influente.

No Apêndice D podem ser consultados dados que completam a análise deste modelo.



(a) (b)

Figura 3.11: Gráficos dos preços observados e preços preditos pelo modelo GAMA INFERIOR.

Modelo Gama alta

Dos modelos construídos para imóveis de gama alta, aquele que produziu melhores resultados foi o construído partir do conjunto de variáveis independentes C3. O método *stepwise* selecionou as variáveis apresentadas na Tabela 3.16 como preditores significativos do preço médio desta gama de imóveis.

Tabela 3.16: Tabela sumária do modelo gama alta.

	$\widehat{\beta}_j$	Erro padrão	$\widehat{\beta}'_j$	t	p
(Constante)	126889,3	160118,1		0,793	0,429
ÁreaÚtil	2343,9	124,5	0,620	18,820	<0,001
Outside	137113,0	23928,4	0,203	5,730	<0,001
NrWCs	77415,9	19681,7	0,124	3,933	<0,001
Inside	-53540,4	11180,8	-0,173	-4,789	<0,001
TempoPub	-35358,1	10152,9	-0,104	-3,483	0,001
TempoCQ	31317,8	9769,7	0,093	3,206	0,001
Estacionamento	73716,6	41776,4	0,067	1,765	0,078
Estado	160520,7	46415,3	0,112	3,458	0,001
Natureza	-246631,8	62508,9	-0,151	-3,945	<0,001

As variáveis com maior contribuição relativa para a explicação do comportamento médio do preço destes imóveis são a **ÁreaÚtil**, **Outside** e **Inside**, seguindo-se **Natureza**, **NrWCs** e **Estado**. Com menos relevância tem-se **TempoPub**, **TempoCQ** e **Estacionamento**.

O modelo final ajustado é:

$$\begin{aligned}\widehat{Preço} =& 126889,3 + 2343,9\boldsymbol{ÁreaÚtil} + 137113,0\boldsymbol{Outside} \\ &+ 77415,9\boldsymbol{NrWCs} - 53540,4\boldsymbol{Inside} - 35358,1\boldsymbol{TempoPub} \\ &+ 31317,8\boldsymbol{TempoCQ} + 73716,6\boldsymbol{Estacionamento} \\ &+ 160520,7\boldsymbol{Estado} - 246631,8\boldsymbol{Natureza} \end{aligned} \tag{3.5}$$

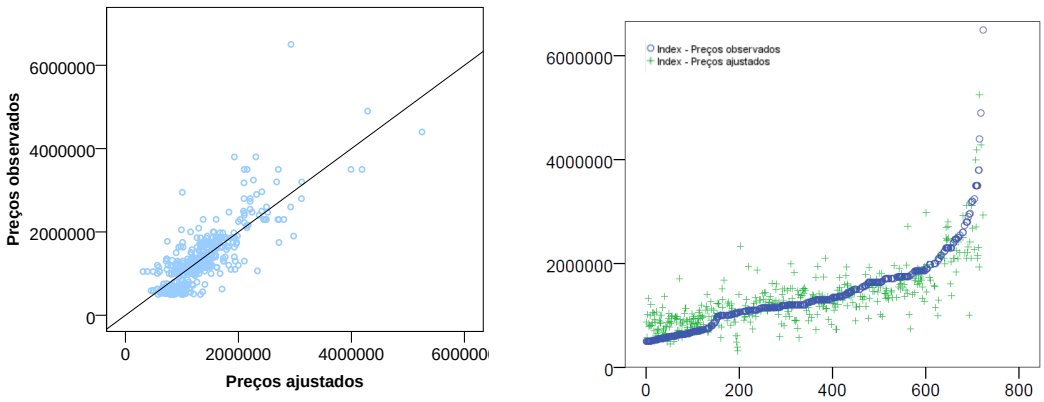
Estas variáveis explicam cerca de 64.8% da variabilidade do preço dos imóveis de gama alta, em torno da sua média ($F \approx 92,005$, $R^2_a \approx 0,648$, $p < 0,001$). A REQM para este modelo é igual a 424607.3.

Os pressupostos do modelo não foram verificados na íntegra, no entanto não foram encontrados problemas de multicolinearidade entre as variáveis envolvidas no modelo (veja-se Apêndice C).

As Figuras 3.12a e 3.12b mostram que as previsões acompanham a tendência dos valores observados, no entanto existe uma variabilidade muito grande.

Foram analisados os *outliers* do modelo não tendo sido observados valores muito discrepantes dos valores teóricos. Também não foram observadas observações com distância de Cook superior a 1, pelo que nenhuma observação foi considerada excessivamente influente.

No Apêndice D podem ser consultados dados que completam a análise deste modelo.



(a) (b)

Figura 3.12: Gráficos dos preços observados e preços preditos pelo modelo GAMA ALTA.

Comparação dos modelos

Tabela 3.17: Resultados globais dos modelos obtidos.

	GERAL	APARTAMENTOS	MORADIAS	GAMA MÉDIA/BAIXA	GAMA ALTA
Tamanho da amostra	1070	937	138	624	446
Variáveis iniciais	44	43	21	44	27
Variáveis selecionadas	11	9	5	18	9
Atributos estruturais selecionados	9	7	4	13	7
Atributos de localização selecionados	2	2	1	5	2
Capacidade explicativa	81,5%	80,6%	65,4%	73,8%	64,8%
REQM	310354,5	238034,7	594309,5	59939,1	424607,3

Nesta análise verificou-se que todos os modelos obtidos selecionam quer atributos estruturais quer atributos de localização o que reforça a ideia de que ambos são importantes para a formação do preço de um imóvel. Pode observar-se ainda que a capacidade explicativa varia entre os 64,8% e os 81,5%, valores que são considerados bons já que, a este nível de especificidade, as formulações dos modelos tornam-se mais sensíveis, especialmente porque são utilizadas muitas variáveis *dummy* e variáveis com relação linear fraca com a variável **Preço** (como os tempos de viagem às categorias de pontos de interesse). Verifica-se também que o número de variáveis selecionadas é bastante reduzido em relação ao número de variáveis iniciais, no entanto, a capacidade explicativa não é muito afetada por esse facto. Contudo verifica-se que os modelos que utilizaram amostras maiores apresentam uma capacidade explicativa maior.

Os primeiros três modelos apresentam valores para a REQM considerados bastante elevados, dada a gama de preços que estes imóveis avaliam. Com a divisão dos imóveis em gama média/baixa e gama alta foi possível reduzir a REQM consideravelmente para o primeiro. E, apesar da REQM ser ainda elevada para o modelo GAMA ALTA, apresenta um valor aceitável dado o intervalo de valores em que o preço desses imóveis se distribui. Embora a divisão dos modelos tenha reduzido o REQM para os modelos APARTAMENTOS e GAMA MÉDIA/BAIXA, a capacidade explicativa dos modelos é menor quando comparada com o modelo GERAL.

Nenhum modelo cumpre na íntegra os pressupostos do modelo de regressão linear múltipla. Esta ocorrência pode estar relacionada com o facto dos dados serem altamente dispersos e das limitações da amostra referidas na secção 3.1.

Foi possível verificar que os atributos estruturais, nomeadamente os gerais, têm um impacto muito grande na formação do preço de um imóvel, com especial destaque para a área útil, que foi selecionada em todos os modelos com a maior contribuição relativa em todos eles. Para além da área útil, o número de casas de banho é também selecionado em todos os modelos com coeficiente positivo e com uma contribuição relativa importante. O estado de conservação é selecionado em todos os modelos, exceto no modelo MORADIAS, com coeficiente de sinal positivo. É também um atributo importante especialmente para o modelo GAMA MÉDIA/BAIXA. A natureza do imóvel não integra os modelos APARTAMENTOS e MORADIAS (como é lógico), no entanto é selecionado no modelo GERAL e no modelo GAMA ALTA, indicando que em média uma moradia é mais cara do que um apartamento, característica que não parece ser importante para os imóveis de gama média/baixa. Para finalizar a análise dos atributos estruturais gerais resta analisar o número de quartos. Este atributo é valorizado apenas nas moradias e nos imóveis de gama média/baixa; era expectável que este atributo fosse importante na formação do preço dos imóveis, no entanto não é selecionado em todos os modelos. O facto desta variável estar muito correlacionada com a área útil pode ser uma explicação para a esta ocorrência.

Relativamente aos atributos estruturais descritivos verifica-se que em geral são atributos que acrescentam valor aos imóveis, apesar de se observar que a existência de elevador, no modelo GERAL, e a existência de quintal, no modelo GAMA MÉDIA/BAIXA, retiram valor ao imóvel. A primeira não apresenta um comportamento expectável já que a existência de elevador é uma facilidade útil para quem vive em apartamentos num andar superior ao rés-do-chão, a segunda pode estar relacionada com a falta de tempo para cuidar deste tipo de espaços.

Relativamente aos atributos de localização, é notório nesta análise que o número de bairros sociais por freguesia é uma variável importante, já que foi selecionada, com coeficiente negativo, por três dos cinco modelos explorados, indicando que a proximidade a bairros sociais diminui o preço de um imóvel. Este comportamento é expectável uma vez que estes locais são apontados como potencialmente problemáticos. Verifica-se ainda que a proximidade a locais de diversão noturna acrescenta valor ao imóvel, tal como o distanciamento a aterros sanitários. Para além destes atributos, não se verifica uma seleção de atributos relacionados com a localização de uma forma consistente nos modelos estudados. Apesar do fator de localização ser apontado como um dos mais importantes é também apontado por Tarré (2009) como o mais difícil de introduzir nos modelos de regressão dada a dificuldade da sua caracterização.

Comparação dos modelos APARTAMENTOS**e** MORADIAS

Como sugere a análise descritiva, as características importantes na formação do preço dos apartamentos e moradias são diferentes. Verifica-se a partir da análise de regressão que ambos

os modelos selecionam a área útil e o número de casas de banho com impacto semelhante na formação do preço dos dois tipos de imóveis. No entanto é possível observar que a **Natureza** é uma variável relevante na formação do preço dos apartamentos que não é nas moradias. Pelo contrário o número de quartos é selecionado no modelo MORADIAS mas não é para o modelo APARTAMENTOS.

Quanto aos atributos estruturais descritivos, apesar do conjunto de atributos selecionado ser diferente, é possível verificar que, em geral, estes atributos acrescentam valor aos dois tipos de imóveis, já que as variáveis são selecionadas com coeficiente positivo. Verifica-se ainda que o impacto destes atributos é semelhante na formação do preço de ambos os tipos de imóvel. Nos apartamentos é possível discriminar que os atributos descritivos com impacto na formação do seu preço são a existência de jardim, piscina, vigilância e vista de rio.

Quanto aos atributos de localização verificam-se algumas diferenças: no modelo APARTAMENTOS é selecionado o número de bairros sociais por freguesia e o tempo ao aterro. Já no modelo MORADIAS apenas é selecionado o tempo ao metro, no entanto, através da análise dos coeficientes *standardizados*, estes atributos têm maior impacto na formação do preço das moradias.

Comparação dos modelos GAMA MÉDIA/BAIXA**e** GAMA ALTA

Na análise de regressão verifica-se, tal como sugere a análise descritiva, que a formação do preço dos imóveis de luxo é diferente dos restantes imóveis. Ambos selecionam a área útil, no entanto esta tem maior impacto na formação do preço dos primeiros, tal como o número de casas de banho. Também o **Estado** é selecionado em ambos os modelos, mas com maior impacto na formação do preço dos imóveis de gama média/baixa. Verifica-se que a natureza do imóvel só tem impacto na formação do preço dos imóveis de gama alta, ao contrário do número de quartos que é relevante apenas para os imóveis de gama média/baixa.

Quanto aos atributos estruturais descritivos existem também algumas diferenças: os atributos relacionados com características do grupo **Inside** têm impacto negativo nos imóveis de gama alta e impacto positivo nos restantes. Os atributos relacionados com características do grupo **Outside** têm impacto positivo nos imóveis de gama alta e negativo nos restantes. A existência de garagem e/ou lugar de estacionamento têm impacto positivo em ambos os modelos, contudo é uma característica mais relevante para os imóveis de gama média/baixa. O modelo GAMA ALTA não seleciona atributos relacionados com a segurança do imóvel, ao contrário do modelo GAMA MÉDIA/BAIXA para os quais é relevante se o imóvel está inserido num condomínio fechado.

Quanto aos atributos de localização, para os imóveis de gama média/baixa o afastamento a bairros sociais, paragens de autocarro, hospitais e aterros valorizam o imóvel, enquanto que para os imóveis de gama alta nenhuma dessas características é relevante. O afastamento ao complexo químico de Algés é uma característica importante apenas para os imóveis de gama alta. Para ambos os tipos de imóveis, a proximidade a locais de diversão noturna é uma característica relevante, contudo com maior impacto na formação do preço dos imóveis de

gama alta.

Capítulo 4

Conclusões e sugestões para trabalho futuro

A reflexão teórica inicial permitiu aprofundar o estudo do tema da habitação, conhecer as características apontadas como as mais relevantes na formação do preço de imóveis e os métodos utilizados na obtenção de um modelo de previsão em trabalhos anteriores, tendo-se revelado importante para o planeamento das características a recolher e para a escolha do método adequado para a obtenção do modelo de previsão.

É importante conhecer os diferentes métodos de avaliação imobiliária e a sua aderência ao mercado em estudo. O método comparativo é o mais utilizado, no entanto para que seja bem implementado é necessário uma base de dados capaz de corresponder às suas exigências e necessita de experiência no processo de ajustamento dos preços. O método do custo é um método muito difundido no mercado imobiliário pois tem por base os custos de produção do bem, no entanto muitos avaliadores entendem que devem ser incorporados fatores como a localização de forma a avaliar corretamente um imóvel. O método de avaliação residual é considerado o mais científico na sua metodologia pois tem por base a gestão urbanística e do planeamento permitindo avaliar o que de facto "está no papel". Os modelos hedónicos são mais recentes e pretendem interpretar as preferências dos potenciais compradores e identificar os atributos que explicam a variabilidade do preço dos imóveis e a importância relativa de cada um deles.

Quanto às características recolhidas, apesar de ter sido estudado que muitas outras são importantes para a formação do preço dos imóveis, houve limitações impostas pela forma como os dados foram recolhidos, pelo que se considera que outras características poderiam ter sido analisadas se os dados fossem recolhidos de outra forma. Para além do estudo das características apontadas como relevantes em trabalhos anteriores era interessante articular esse estudo com a elaboração de um inquérito aos indivíduos que procuram comprar casa em Lisboa na tentativa de encontrar outras características que ainda não foram estudadas mas que na realidade importam na altura de escolher o imóvel a comprar, principalmente as características valorizadas pelos habitantes ou futuros habitantes da área em estudo. As

limitações impostas pela escassez de dados pode traduzir-se numa recolha insuficiente de atributos ou utilização de atributos que não são efetivamente importantes.

Quanto ao método de seleção de variáveis e obtenção do modelo de previsão, foi selecionado o modelo de regressão linear múltipla com seleção *stepwise* sem transformações. Esta seleção baseou-se no facto de não se pretender transformar as variáveis, de forma a ser possível analisar a contribuição de cada atributo no modelo; a seleção *stepwise* foi escolhida por ser um método que seleciona variáveis parcimoniosamente e evita problemas de multicolinearidade. O ponto forte da aplicação do método de regressão linear múltipla é que a sua análise permite especificar as variáveis relevantes na formação do preço dos imóveis e quando executada com êxito produz estimativas da contribuição relativa de cada uma dessas variáveis. No entanto apresenta algumas limitações: é uma técnica exigente em recursos, nomeadamente porque requer informação quantitativa relativamente a um número elevado de imóveis; a recolha da informação pode ser demorada, dispendiosa ou mesmo pouco exequível; a análise de regressão pode chegar à conclusão de que existe uma forte ligação entre duas variáveis, não levando em conta a influência potencial de outras variáveis que podem ser ainda mais importantes, e que se correlacionam com as estudadas; as relações entre as diferentes variáveis explicadas e explicativas são, muitas vezes, cíclicas (X explica Y e Y explica X); as observações devem apresentar situações com variância suficiente para permitir o ajustamento (por exemplo, se todas as observações disserem respeito a apartamentos não será possível estimar a influência da natureza do imóvel na formação dos preços (Costa, 2009)). A ferramenta deve ser, assim, usada com precaução.

Apesar das limitações descritas foi possível obter modelos com uma capacidade explicativa bastante boa e com um número de variáveis reduzido. Como pretendido, foi possível descrever o preço dos imóveis de habitação no concelho de Lisboa utilizando atributos quer estruturais quer de localização. Foi possível identificar que o comportamento dos preços de um apartamento e de uma moradia é diferente pelo que foram apresentados modelos diferentes para estes tipos de imóveis. Foi ainda identificado que o comportamento dos preços para imóveis de gama alta é diferente dos restantes.

Os principais aspetos a serem melhorados são:

- Determinação de uma forma mais eficiente de classificar os atributos de localização de um imóvel. Existem trabalhos que abordam a classificação destes atributos de outra forma, nomeadamente atribuindo uma pontuação ao local em que está situado o imóvel conforme as acessibilidades que o rodeiam. Essa alternativa foi ponderada mas não foi possível implementar devido à forma como foram recolhidos os dados relativos à localização do imóvel.

- Definir uma boa estratégia de recolha de dados e construção fundamentada de novos atributos. Elaborar inquéritos de forma a identificar atributos importantes na formação do preço dos imóveis que não foram considerados ou recolher informação de uma forma

mais eficiente, de forma a garantir uma análise simplificada e modelos aplicáveis.

- Conhecer o local em estudo. Fazer visitas, identificar locais importantes na formação do preço dos imóveis, articular os conhecimentos de uma pessoa de ordenamento de território de forma a classificar a localização de forma mais eficiente.

- Conhecer melhor o mercado em estudo. Articular conhecimentos com profissionais do mercado imobiliário da área territorial em estudo. Os profissionais imobiliários, pela sua proximidade ao mercado, têm um maior conhecimento sobre o comportamento da oferta e da procura, dos preços, das tendências e das flutuações do mercado.

O trabalho permitiu ainda aprofundar os conhecimentos do *software* utilizado, o SPSS IBM *Statistics*, nomeadamente na área da regressão linear e testes de hipóteses.

Capítulo 5

Bibliografia

Tavares, F. A. O. (2011). «Avaliação Imobiliária - Entre a Ciência da Avaliação e a Arte da Apreciação». PhD. Universidade de Aveiro.

González, M. e C. Formoso (2000). «Análise conceitual das dificuldades na determinação de modelos de formação de preços através de análise de regressão». Em:

Batista, P.R.L. (2010). «Data mining na identificação de atributos valorativos da habitação». MSc. Universidade de Aveiro.

Catalão, A.T.M. (2010). «Estudo do Mercado Imobiliário de Aveiro». MSc. Universidade de Aveiro.

Guedes, T.B. (2011). «Polinómios fracionários na modelação do preço de imóveis». MSc. Universidade de Aveiro.

Marques, J.J.L. (2012). «A noção de espaço nos mesrcados habitacionais urbanos». PhD. Universidade de Aveiro.

Vigas, M.L.B. (2013). «Índice de preços imobiliários: um exercício na área Aveiro-Ílhavo». MSc. Universidade de Aveiro.

Couto, P. (2007). «Avaliação patrimonial de imóveis para habitação». PhD. Universidade do Porto.

Moreira, D.S., R.S. Silva e A.M.R. Fernandes (2010). «Engenharia de avaliações de imóveis apoiada em técnicas dde análise multicritério e redes neurais artificiais». Em: *Revista de Sistemas de Informação do FSMA* 6, pp. 49–58.

Kiel, K. A. e J. E. Zabel (2008). «Location, location, location: The 3L Approach to house price determination». Em: *Journal of Housing Economics* 17.2, pp. 175–190.

Montgomery, D.C., E.A. Peck e G.G. Vining (2006). *Introduction to linear regression analysis*. 4ª Ed. Wiley Series in Probability and Statistics. EUA: Wiley.

Murteira, B. et al. (2010). *Introdução à Estatística*. Lisboa: Escolar Editora.

Marôco, João (2010). *Análise Estatística com o PASW Statistics*. Ed. por Lda ReportNumber. Pêro Pinheiro. 953 pp.

Field, A. (2009). *Discovering Statistics using SPSS*. SAGE Publications.

Hall, A., C. Neves e A. Pereira (2011). *Grande maratona de estatistica no SPSS*. Escolar Editora.

Snee, R.D. (1997). «Validation of Regression Models: Methods and Examples». Em: *Technometrics*. Taylor & Francis Group 19.4, pp. 415–428.

Tabachnick, Barbara G. e Linda S. Fidell (2007). *Using Multivariate Statistics*. Ed. por Pearson Education. 3ª Ed.

Ribeiro, Rita. *Preço das casas de luxo em Lisboa aumenta 54% influenciado pelos Golden Visa*. type. Lisboa: Confidencial Imobiliário. URL: http://www.confidencialimobiliario.com/?q=content/press-release-preco-das-casas-de-luxo-em-lisboa-aumenta-54-influenciado-pelos-golden-visa.

Neto, F. (2008). «Aplicação de um modelo hedónico de avaliação a edifícios habitacionais no concelho de Gaia». MSc. Universidade Técnica de Lisboa.

Tarré, Ana (2009). «Análise de valores de avaliação de apartamentos no âmbito do Crédito a Habitação, para duas zonas distintas do concelho de Lisboa - recurso a Modelos Hedónicos». MSc. Lisboa: Universidade Técnica de Lisboa.

Costa, C. (2009). «A Avaliação do Desenvolvimento Socioeconómico, MANUAL TÉCNICO II: Métodos e Técnicas». Em: *Agência para o desenvolvimento e coesão*.

Apêndice A

Estatísticas Descritivas

Neste anexo encontram-se algumas estatísticas descritivas que servem de apoio à análise descritiva do problema que é efetuada na secção 2.2.

Tabela A.1: Estatísticas descritivas das variáveis quantitativas contínuas da amostra.

		TempoAutocarro	TempoCafé	TempoMercearia	TempoGinásio	TempoHospital	TempoPub	TempoParque
N	Valid	1471	1471	1471	1471	1471	1471	1471
	Missing	0	0	0	0	0	0	0
Mean		1,354	1,047	1,370	2,022	2,412	3,266	2,211
Median		1,000	1,000	1,000	2,000	2,000	3,000	2,000
SD		1,278	1,404	1,475	1,505	1,742	2,088	1,619
Skewness		0,962	1,642	2,056	1,176	0,890	0,592	0,826
SD Skewness		0,064	0,064	0,064	0,064	0,064	0,064	0,064
Kurtosis		,711	2,218	5,166	2,467	,892	-,284	,361
SE Kurtosis		,128	,128	,128	,128	,128	,128	,128
Range		5,300	6,000	7,000	8,333	9,567	11,033	10,183
Minimum		0,000	0,000	0,000	0,000	0,000	0,000	0,000
Maximum		5,300	6,000	7,000	8,333	9,567	11,033	10,183

		TempFarmácia	TempoCQ	TempoRestaurante	TempoEscola	TempoMetro	TempoAterro	TempoZI
N	Valid	1471	1471	1471	1471	1471	1471	1471
	Missing	0	0	0	0	0	0	0
Mean		1,472	11,736	1,050	1,309	3,353	11,129	8,054
Median		1,000	11,000	1,000	1,000	3,000	11,133	8,000
Std. Deviation		1,381	2,512	1,342	1,015	2,616	2,564	2,144
Skewness		1,277	0,069	1,592	0,906	0,759	-0,304	-0,693
SD Skewness		0,064	0,064	0,064	0,064	0,064	0,064	0,064
Kurtosis		1,631	-,162	1,886	,962	-,347	-,435	,443
SE Kurtosis		,128	,128	,128	,128	,128	,128	,128
Range		6,000	13,883	5,267	6,467	11,000	15,150	11,900
Minimum		0,000	6,000	0,000	0,000	0,000	0,900	2,000
Maximum		6,000	19,883	5,267	6,467	11,000	16,050	13,900

Apêndice B

Validação dos modelos

Como foi explicado no relatório, para cada abordagem, foram comparados modelos construídos a partir de três subconjuntos de atributos, o que totaliza 15 modelos estudados. Cada um dos 15 modelos passou pela técnica de validação cruzada, de forma a selecionar o "melhor" para cada abordagem, resultando assim nos cinco modelos analisados no Capítulo 2.

Para cada subconjunto de variáveis e para cada abordagem, foi aplicado o seguinte procedimento cinco vezes:

1. Dividiu-se a amostra em conjunto de treino (80 % das observações) e conjunto de teste (20 % das observações) de forma aleatória.
2. Aplicou-se o modelo de regressão linear múltipla com seleção de variáveis *stepwise* ao conjunto de treino obtido no ponto anterior.
3. O modelo obtido no ponto 1. foi aplicado a dez subconjuntos de treino.
4. Foram determinados os valores de R^2_{treino} e R^2_{teste}.
5. Determinação da média dos valores de R^2_{treino} e R^2_{teste} obtidos nos pontos anteriores.

Após o procedimento anterior, determinou-se a média dos valores de R^2_{treino} e R^2_{teste} obtidos. Os resultados são apresentados na Tabela B.1.

Foram selecionados os modelos com valores médios de R^2_{treino} e R^2_{teste} maiores e menores diferença entre os mesmos, um para cada abordagem e subconjunto de variáveis independentes da Tabela 3.8 num total de 15 modelos. De seguida, a partir do mesmo critério de avaliação, foi selecionado um modelo para cada abordagem, o qual ditou as variáveis independentes a utilizar na construção dos modelos finais. Estes últimos modelos foram construídos utilizando a amostra completa em cada abordagem. Os resultados podem ser consultados na discussão da Secção 2.3.

Tabela B.1: Média dos valores de R^2_{treino} e R^2_{teste} obtida na aplicação da técnica de validação cruzada

R^2		Treino	Teste	Treino	Teste	Treino	Teste	Treino	Teste	Treino	Teste
GERAL	C1	0,819	0,804	0,819	0,737	0,823	0,813	0,811	0,818	0,817	0,802
	C2	0,799	0,777	0,795	0,794	0,794	0,787	0,801	0,769	0,791	0,810
	C3	0,809	0,779	0,806	0,795	0,808	0,795	0,808	0,779	0,802	0,814
APARTAMENTOS	C1	0,806	0,814	0,807	0,800	0,805	0,814	0,806	0,817	0,811	0,799
	C2	0,800	0,764	0,789	0,803	0,796	0,766	0,789	0,811	0,787	0,796
	C3	0,798	0,796	0,795	0,806	0,793	0,799	0,800	0,797	0,801	0,779
MORADIAS	C1	0,697	0,658	0,642	0,523	0,627	0,613	0,631	0,595	0,659	0,613
	C2	0,675	0,623	0,697	0,674	0,629	0,659	0,677	0,662	0,668	0,666
	C3	0,683	0,642	0,704	0,666	0,683	0,706	0,650	0,590	0,666	0,693
GAMA MÉDIA/BAIXA	C1	0,743	0,715	0,746	0,737	0,741	0,728	0,739	0,700	0,741	0,722
	C2	0,726	0,705	0,727	0,722	0,715	0,718	0,430	0,376	0,727	0,716
	C3	0,740	0,726	0,735	0,709	0,642	0,667	0,722	0,731	0,651	0,642
GAMA ALTA	C1	0,649	0,642	0,634	0,652	0,661	0,645	0,664	0,651	0,669	0,651
	C2	0,634	0,604	0,615	0,660	0,595	0,620	0,618	0,661	0,626	0,600
	C3	0,664	0,641	0,644	0,643	0,653	0,656	0,622	0,641	0,612	0,595

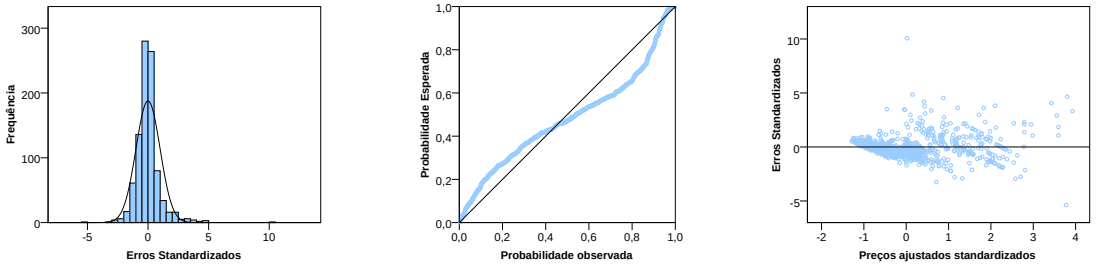
Apêndice C

Verificação de pressupostos dos modelos de regressão

C.1 Modelo Apartamentos

O pressuposto da independência dos erros foi validado com a estatística de teste de Durbin-Watson ($d \approx 1,129$). O pressuposto da normalidade dos erros foi analisado graficamente (ver Figura C.1), não tendo sido validado, pois os valores obtidos distam consideravelmente dos valores de referência. No entanto Murteira et al. (2010) referem que na presença de grandes amostras o facto do pressuposto da normalidade dos erros não ser validado não anula a inferência sobre o modelo de regressão desde que os erros tenham média nula e sejam homocedásticos. Como por definição os erros têm valor médio nulo, basta verificar se estes são homocedásticos. A fim de validar esse pressuposto foi construído o gráfico da Figura C.1c de onde se conclui que os erros não têm variância constante, conclusão que é reforçada pelo resultado do teste de White-simplificado ($W_s \approx 59,030 > \chi(0.95,2) \approx 5,991$), nestas circunstâncias o estimador dos mínimos quadrados continua a ser consistente e não enviesado, no entanto deixa de ser o mais eficiente entre os estimadores lineares não enviesados. Contudo, para Tabachnick e Fidell (2007) o facto de os erros não serem homocedásticos não invalida a análise de regressão, apenas a enfraquece. Recorde-se que o estimador dos mínimos quadrados, nesta circunstância, continua a ser consistente e não enviesado, simplesmente deixa de ser o mais eficiente entre os estimadores lineares não enviesados.

Da análise da Tabela C.1, verifica-se que não existem problemas de colinearidade entre as variáveis envolvidas no modelo, dado que os valores de VIF são inferiores a 10, sendo a sua média $1,490$, e os valores de *tolerance* superiores a $0,2$.



(a) (b) (c)

Figura C.1: Gráficos dos erros de regressão do modelo APARTAMENTOS.

Tabela C.1: Valores de *tolerance* e VIF das variáveis selecionadas pelo modelo APARTAMENTOS.

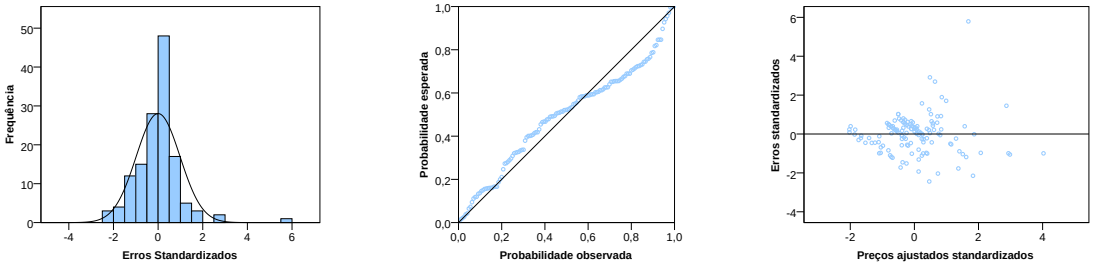
	Tolerance	VIF
ÁreaÚtil	0,395	2,533
Estado	0,857	1,167
Piscina	0,767	1,304
NrWCs	0,374	2,675
TempoAterro	0,900	1,111
Vigilância	0,887	1,128
Jardim	0,784	1,275
NrBairrosSociais	0,907	1,102
VistaRio	0,894	1,118

C.2 Modelo MORADIAS

O pressuposto da independência dos erros foi validado com a estatística de teste de Durbin-Watson ($d \approx 1,152$). O pressuposto da normalidade dos erros não foi verificado. De forma a verificar se os erros são homocedásticos foi construído o grágico da Figura C.2c e efetuado o teste de White-simplificado, para o qual se obteve $W_s \approx 6,486 > \chi(0.95, 2) \approx 5,991$, a partir dos quais se verifica-se que os erros não são homocedásticos.
Da análise da Tabela C.2, verifica-se que não existem problemas de colinearidade entre as variáveis envolvidas no modelo, dado que os valores de VIF são inferiores a 10, sendo a sua média $1,405$, e os valores de *tolerance* superiores a $0,2$.

Tabela C.2: Valores de *tolerance* e VIF das variáveis selecionadas pelo modelo MORADIAS.

	Tolerance	VIF
ÁreaÚtil	0,590	1,693
TempoMetro	0,858	1,166
NrWCs	0,817	1,224
NrCaracterísticas	0,916	1,092
NrQuartos	0,541	1,848



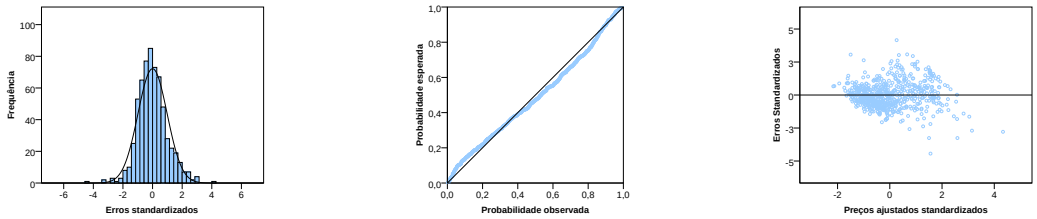
(a) (b) (c)

Figura C.2: Gráfico dos erros de regressão do modelo MORADIAS.

C.3 Modelo GAMA MÉDIA/BAIXA

O pressuposto da independência dos erros foi validado com a estatística de teste de Durbin-Watson ($d \approx 1,358$). O pressuposto da normalidade dos erros não foi verificado. De forma a verificar se os erros são homocedásticos foi construído o grágico da Figura C.2c e efetuado o teste de White-simplificado, para o qual se obteve $W_s \approx 48,672 > \chi(0.95,2) \approx 5,991$, a partir dos quais se verifica-se que os erros não são homocedásticos.

Da análise da Tabela C.3, verifica-se que não existem problemas de colinearidade entre as variáveis envolvidas no modelo, dado que os valores de VIF são inferiores a 10, sendo a sua média $1,635$, e os valores de *tolerance* superiores a $0,2$.



(a) (b) (c)

Figura C.3: Gráfico dos erros de regressão do modelo GAMA MÉDIA/BAIXA.

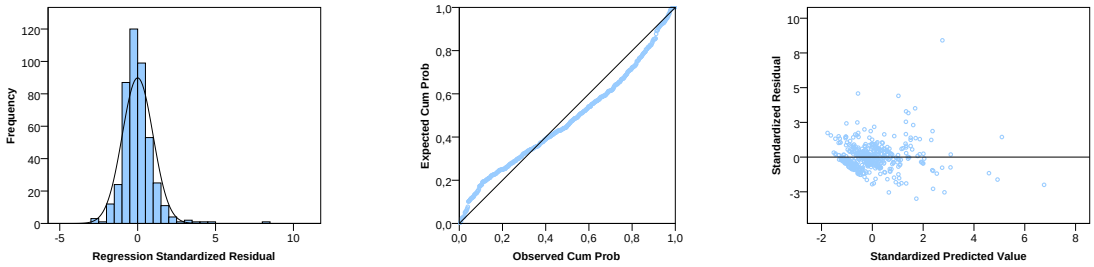
Tabela C.3: Valores de *tolerance* e VIF das variáveis selecionadas pelo modelo GAMA MÉDIA/BAIXA.

	Tolerance	VIF
Área útil	0,308	3,250
Estado	0,727	1,375
Aquecimento Central	0,750	1,333
Elevador	0,622	1,607
Número de bairros sociais por freguesia	0,642	1,558
Tempo ao pub	0,654	1,529
Ar Condicionado	0,738	1,355
Nº de Quartos	0,389	2,570
Condomínio Fechado	0,806	1,240
Estacionamento	0,620	1,614
Quintal	0,977	1,024
Garagem	0,849	1,178
Nº de WCs	0,433	2,307
Tempo ao aterro	0,638	1,568
Tempo ao hospital	0,504	1,983
Tempo ao autocarro	0,685	1,461
Cozinha Equipada	0,726	1,377
Piscina	0,902	1,108

C.4 Modelo GAMA ALTA

O pressuposto da independência dos erros foi validado com a estatística de teste de Durbin-Watson ($d \approx 1,358$). O pressuposto da normalidade dos erros não foi verificado. De forma a verificar se os erros são homocedásticos foi construído o grágico da Figura C.2c e efetuado o teste de White-simplificado, para o qual se obteve $W_s \approx 6,486 > \chi(0.95, 2) \approx 5,991$, a partir dos quais se verifica-se que os erros não são homocedásticos.

Da análise da Tabela C.4, verifica-se que não existem problemas de colinearidade entre as variáveis envolvidas no modelo, dado que os valores de VIF são inferiores a 10, sendo a sua média $1,450$, e os valores de *tolerance* superiores a $0,2$.



(a) (b) (c)

Figura C.4: Gráfico dos erros de regressão do modelo GAMA ALTA.

Tabela C.4: Valores de *tolerance* e VIF das variáveis selecionadas pelo modelo GAMA ALTA.

	Tolerance	VIF
Área útil	0,730	1,370
Outside	0,627	1,595
NrWCs	0,798	1,254
Inside	0,604	1,656
TempoPub	0,893	1,119
TempoCQ	0,932	1,073
Estacionamento	0,554	1,804
Estado	0,752	1,330
Natureza	0,541	1,850

Apêndice D

Análise descritiva das variáveis selecionadas pelos modelos

D.1 Modelo Apartamentos

Os preços dos apartamentos da amostra distribuem-se entre os 39500 e os 3500000 euros. É um intervalo de valores aceitáveis, contudo alguns apartamentos apresentam preços bastantes elevados. Também se verifica que o desvio-padrão é de aproximadamente 650 mil euros, que é um valor bastante elevado.

A área útil varia entre 18 m^2 e 2300m^2 e apresenta valores atípicos, uma vez que 2300 m^2 é um valor muito elevado para a área útil de um apartamento.

A variável **Preço** tem uma correlação significativa com todas as variáveis envolvidas no modelo, como se pode verificar da análise da Tabela D.2, sendo que a correlação com a **ÁreaÚtil** e **NrWCs** é forte, o que justifica o facto de estas variáveis terem uma contribuição relativa maior para o modelo do que as restantes variáveis apresentadas na tabela. Pode observar-se que os sinais das correlações são iguais aos sinais dos coeficientes do modelo de regressão.

Verifica-se que o preço dos apartamentos é, para a maioria das observações, mais elevado se estes forem novos, tal como se verifica no modelo de regressão.

Tabela D.1: Estatísticas descritivas de variáveis selecionadas pelo modelo apartamentos.

	Preço	ÁreaÚtil	NrWCs	TempoAterro	NrBairrosSociais
média	679746,936	181,608	2,256	11,088	2,443
desvio-padrão	647922,387	133,869	1,190	2,548	1,681
mediana	384000,000	140,000	2,000	11,067	2,000
minimo	39500,000	18,000	1,000	3,000	0,000
máximo	3500000,000	2300,000	7,000	16,050	11,000

Tabela D.2: Coeficientes de correlação de Pearson entre variáveis do modelo APARTAMENTOS.

	Preço	ÁreaÚtil	NrWCs	TempoAterro
Preço				
ÁreaÚtil	0,772**			
NrWCs	0,747**	0,766**		
TempoAterro	0,091**	0,042	-0,094**	
NrBairrosSociais	-0,286**	-0,198**	-0,164**	-0,201**

**: a correlação é significativa ao nível de 0.01

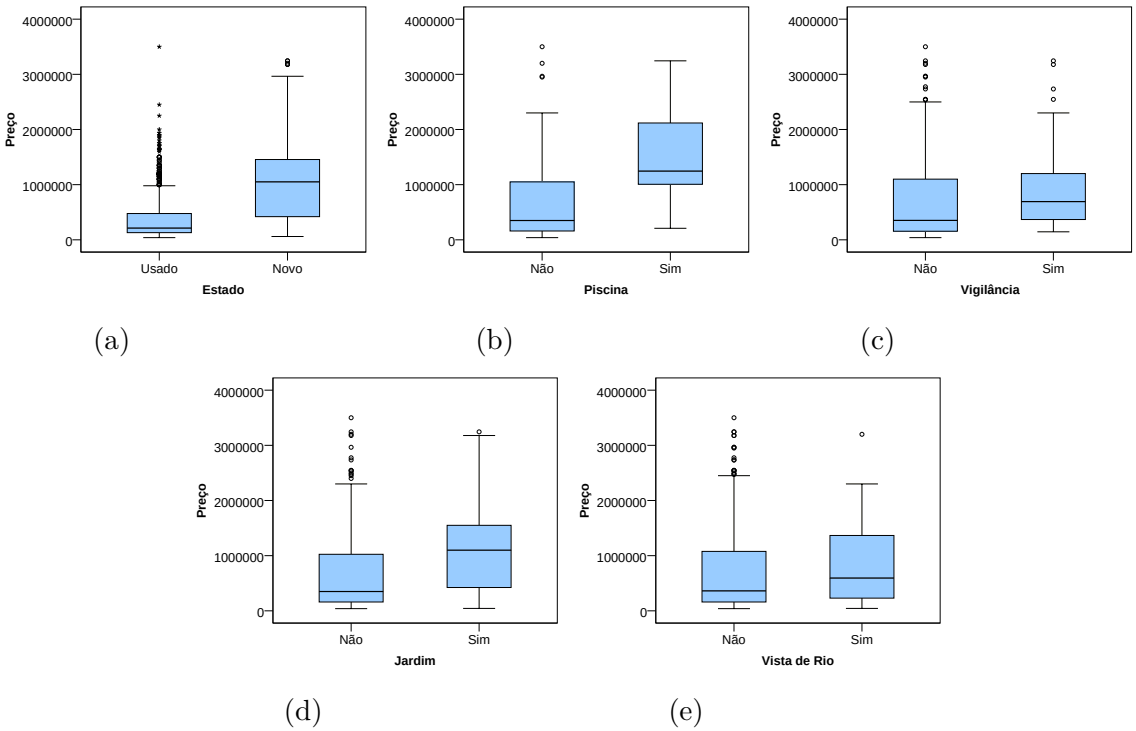


Figura D.1: Box-plots das variáveis (a)**Estado**, (b)**Piscina**, (c)**Vigilância**, (d)**Jardim** e (e)**VistaRio** em função do preço dos apartamentos.

D.2 Modelo MORADIAS

As moradias da amostra apresentam preços entre os 49900 euros e os 6500000 euros. Pelo que existem moradias com preços muito baixos e outras com preços exageradamente elevados. O desvio-padrão desta variável é também bastante elevado, apresentando um valor aproximadamente de 1 milhão de euros. A área útil apresenta um valor mínimo de 40 m^2 que é baixo, dado que é uma moradia, e um valor máximo elevado de 2225 m^2 que é no entanto aceitável. As restantes variáveis envolvidas no modelo apresentam valores plausíveis.

Quantos ás correlações, verifica-se que os sinais são iguais aos dos coeficientes do modelo de regressão. Pode observar-se que o **Preço** apresenta uma correlação forte com a **ÁreaÚtil**, que se verificou ser a variável com maior contribuição relativa no modelo de regressão. Verifica-se

também que as correlações entre o **Preço** e o **TempoAterro** e **NrCaracterísticas** são muito fracas e não significativas, contudo foram selecionadas pelo modelo como preditores significativos.

Tabela D.3: Estatísticas descritivas das variáveis selecionadas pelo modelo MORADIAS.

	Preço	ÁreaÚtil	TempoMetro	NrWCs	NrCaracterísticas	NrQuartos
média	1668889,130	440,880	4,110	3,800	4,870	5,830
desvio-padrão	1010305,814	335,225	3,563	1,382	4,223	2,446
mediana	1747000,000	302,000	4,000	4,000	4,000	5,000
minimo	49900,000	40,000	0,000	1,000	0,000	2,000
máximo	6500000,000	2224,000	11,000	8,000	20,000	10,000

Tabela D.4: Coeficientes de correlação de Pearson entre as variáveis do modelo MORADIAS.

	Preço	ÁreaÚtil	TempoMetro	NrWCs	NrCaracterísticas
Preço					
ÁreaÚtil	0,755**				
TempoMetro	-0,035	0,192**			
NrWCs	0,452**	0,339**	-0,036		
NrCaracterísticas	0,069	-0,094	-0,224**	0,126	
NrQuartos	0,445**	0,573**	0,269**	0,361**	-0,092

**: a correlação é significativa ao nível de 0.01

D.3 Modelo GAMA MÉDIA/BAIXA

Para este modelo, o preço dos imóveis foi limitado até 500 mil euros, pelo que o intervalo em que os preços se distribuem é menor do que nos restantes modelos. Ainda assim o desvio-padrão desta variável é elevado. As restantes variáveis apresentam valores plausíveis.

Verifica-se que todas as variáveis apresentadas na Tabela D.6 apresentam correlações significativas com a variável **Preço**. Esta variável apresenta uma correlação moderada com as variáveis **NrWCs** e **NrQuartos**, no entanto a variável com maior contribuição relativa no modelo de regressão é a **ÁreaÚtil**.

É notório que imóveis em condomínio fechado têm preços mais elevados, nesta gama de imóveis. Relembre-se que esta variável apenas é selecionada no modelo GAMA MÉDIA/BAIXA. Também a existência de piscina surge em imóveis mais caros. Estas conclusões são concordantes com as obtidas da análise do modelo de regressão.

Tabela D.5: Estatísticas descritivas de variáveis selecionadas pelo modelo GAMA MÉDIA/BAIXA.

	média	desvio-padrão	mediana	minimo	máximo
Preço	221783,224	117186,022	190000,000	39500,000	499500,000
ÁreaÚtil	100,618	43,499	95,000	18,000	350,000
NrBairros	2,854	1,848	3,000	0,000	11,000
TempoPub	3,363	2,012	3,000	0,000	11,033
NrQuartos	2,410	1,193	2,000	0,000	10,000
NrWCs	1,658	2,000	0,722	1,000	4,000
TempoAterro	10,818	2,844	11,000	0,900	16,050
TempoHospital	2,756	1,855	3,000	0,000	9,567
TempoAutocarro	1,275	1,423	1,000	0,000	5,300

Tabela D.6: Coeficientes de correlação de Pearson entre variáveis do modelo GAMA MÉDIA/BAIXA.

	Preço	ÁreaÚtil	NrBairros	TempoPub	NrQuartos	NrWCs	TempoAterro	TempoHospital
Preço								
ÁreaÚtil	0,295**							
NrBairros	-0,156**	-0,045						
TempoPub	-0,108**	-,072*	0,066					
NrQuartos	0,511**	0,347**	-0,061	-0,068				
NrWCs	0,662**	0,700**	-0,025	-0,035	0,548**			
TempoAterro	-0,175**	-0,086*	-0,189**	-0,244**	-0,109**	-0,226**		
TempoHospital	-0,180**	-0,115**	0,315**	0,515**	-0,091*	-0,092*	0,051	
TempoAutocarro	0,105**	-0,003	-0,200**	0,334**	0,022	0,063	-0,177**	0,346**

*: a correlação é significativa ao nível de 0.05
**: a correlação é significativa ao nível de 0.01

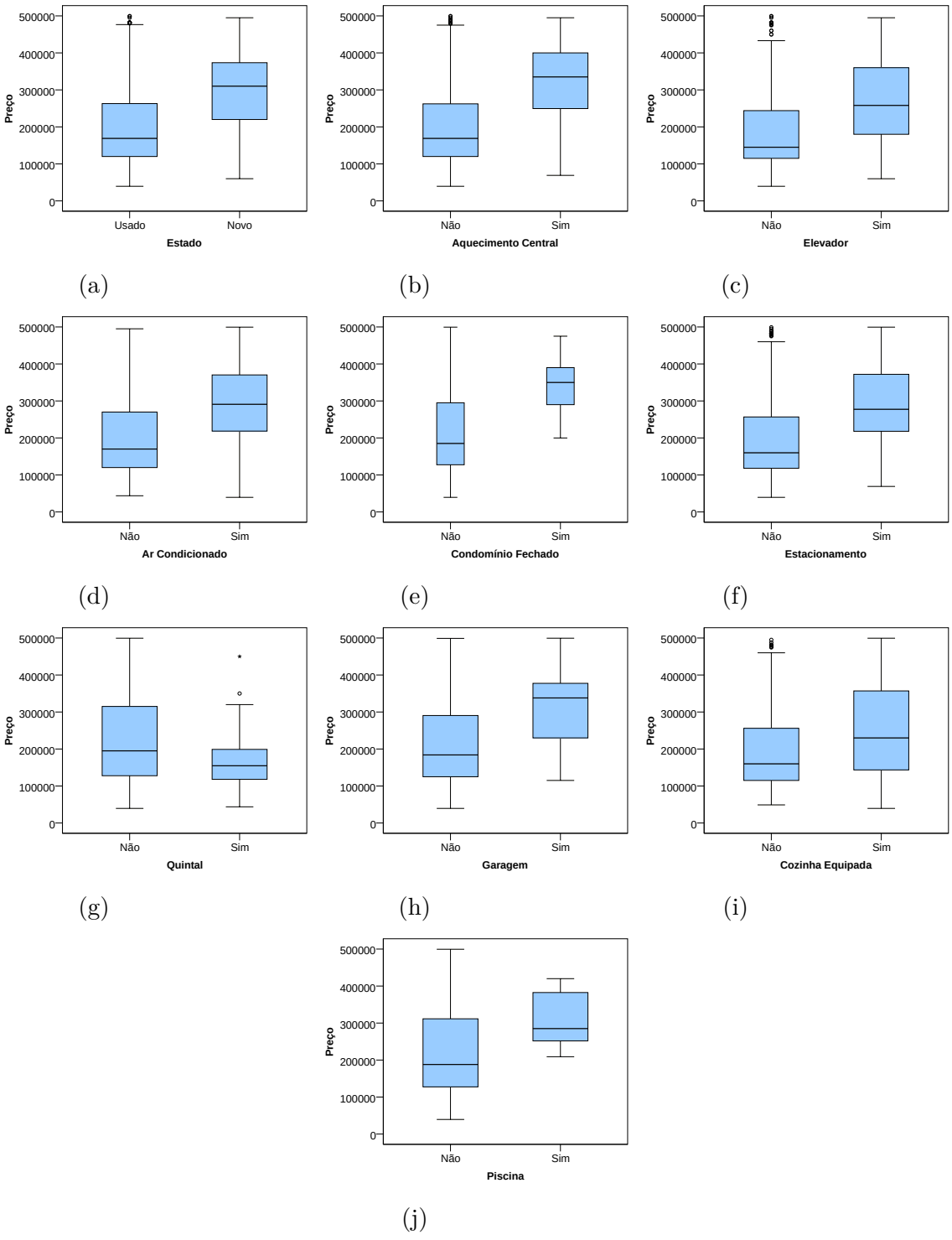


Figura D.2: Box-plots das variáveis (a) **Estado**, (b) **AquecimentoCentral**, (c) **Elevador**, (d) **ArCondicionado**, (e) **CondomínioFechado**, (f) **Estacionamento**, (g) **Quintal**, (h) **Garagem**, (i) **CozinhaEquipada** e (j) **Piscina** em função do preço dos imóveis de gama média/baixa.

D.4 Modelo Gama alta

Para construir este modelo foram selecionados imóveis com preços iguais ou superiores a 500 mil euros, pelo que esse é o valor mínimo da variável **Preço**. Esta variável toma valores até 6 milhões e 500 mil euros. A **ÁreaÚtil** apresenta alguns valores extremos (maiores que 1000 m^2), no entanto são referentes a moradias. As restantes variáveis apresentadas na Tabela D.7 apresentam valores plausíveis.

Verifica-se que nem todas as variáveis do modelo apresentam correlações significativas com a variável **Preço**. Contudo, aquelas que apresentam uma correlação significativa com esta variável são as que têm maior contribuição relativa no modelo de regressão.

O modelo de regressão indica que o imóvel ser novo valoriza-o, no entanto existem imóveis usados com preços bastante superiores. A variável **Estacionamento** foi apontada como significativa na formação do preço deste imóveis, com coeficiente positivo, no entanto verifica-se que os imóveis com preços mais elevados não têm indicação da presença de estacionamento ou lugar de garagem.

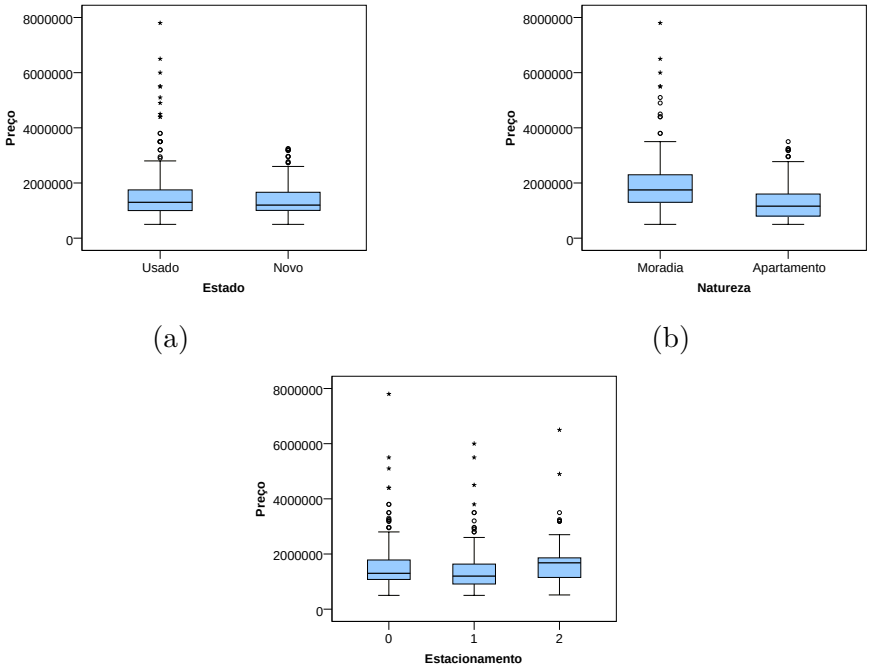
Tabela D.7: Estatísticas descritivas de variáveis selecionadas pelo modelo gama alta.

	Preço	ÁreaÚtil	Outside	NrWCs	Inside	TempoPub	TempoCQ
média	1353099,278	317,930	1,170	3,537	2,747	11,089	3,330
desvio-padrão	730777,513	187,816	1,086	1,170	2,308	2,152	2,148
mediana	1200000,000	263,000	1,000	4,000	3,000	11,000	3,000
minimo	500000,000	59,000	0,000	1,000	0,000	6,000	0,000
máximo	6500000,000	1890,000	5,000	8,000	9,000	18,767	8,000

Tabela D.8: Coeficientes de correlação de Pearson entre variáveis do modelo gama alta.

	Preço	ÁreaÚtil	Outside	NrWCs	Inside	TempoPub
Preço						
ÁreaÚtil	,741**					
Outside	,184**	0,036				
NrWCs	,439**	,375**	,205**			
Inside	-,187**	-,164**	,360**	-0,054		
TempoPub	0,028	,090*	,086*	0,076	-0,068	
TempoCQ	0,007	-,126**	0,019	-0,024	0,062	-,239**

*: a correlação é significativa ao nível de 0.05
**: a correlação é significativa ao nível de 0.01



(a) (b)

(c)

Figura D.3: Box-plots das variáveis (a)**Estado**, (b)**Natureza** e (c)**Estacionamento** em função do preço dos imóveis de gama alta.